Experimental Demonstration of Scalable and Low Latency Crowd Management Enabled by 5G and AI in an Accelerated Edge Cloud

J. C. Borromeo[(1)], K. Kondepu[(2)], S. Fichera[(1)], P. Castoldi[(1)], L. Valcarenghi[(1)]
[(1)]Scuola Superiore Sant'Anna, Pisa, Italy, luca.valcarenghi@santannapisa.it;
[(2)]Indian Institute of Technology Dharwad, Dharwad, India

Abstract: This demo shows how crowd management can benefit from 5G connectivity and accelerated edge cloud AI-based computation to achieve low latency and scalability with respect to a mobile device-based physical distancing computation.
OCIS codes: (060.0060) Fiber optics and optical communications; (060.4250) Networks;

1. Overview

The widespread deployment of edge micro data centers (i.e., edge cloud) [1] together with the support of 5G communication, enables new applications based on data elaboration exploiting Artificial Intelligence/Machine Learning (AI/ML) [2], which can be impractical in end user devices. Indeed, although user equipment (UE) computational capabilities are continuously increasing, such heavy elaborations would compromise their battery duration. For example, video surveillance, search and rescue operation in a hostile environment, traffic management in smart cities, and emergency management by means of Unmanned Aerial Vehicles (UAV) represents realistic scenarios that can benefit from elaboration offloaded to the edge cloud to spare UAV energy [3].
Recently, the COVID-19 pandemic has fostered the development of applications related to crowd management. Social or also called physical distancing is one of the main countermeasures to fight the diffusion of SARS-CoV-2 virus. Applications performing physical distancing verification based on AI/ML are therefore emerging [4]. Often, local police might resort to portable battery powered devices, such as portable cameras or smartphones, to run such applications. Although the computational power of such devices is continuously increasing, physical distancing applications, if locally run in the portable devices could rapidly exhaust battery power and slow them down. This demo shows that offloading the computationally intensive and energy demanding physical distancing application, based on AI/ML, to an accelerated edge cloud, enabled by 5G connectivity, can provide shorter response time and lower energy consumption than the ones achievable by running the application in portable devices.
The proposed demonstration is depicted in Fig. 1. A smartphone acquires the images of an area (e.g., part of the conference premises) where people are present. The raw images are sent to an edge cloud through a 5G connection. The edge cloud is equipped with CPUs, GPUs and FPGAs (i.e., it is accelerated). The edge cloud hosts 5G virtualized Radio Access Network (vRAN) and Next Generation Core (vNGC) network functions. Moreover, it hosts an application for physical distancing computation based on AI/ML. The images conveyed by the 5G fronthaul and backhaul network are elaborated (and also anonymized and not stored for privacy issues) by the physical distancing application and a reply is sent back to the UE, where either a green (i.e., physical distancing fulfilled) or red (i.e., physical distancing not fulfilled) symbol is visualized. Visitors to the demo will experience the capability of the accelerated edge cloud and 5G network to provide a real time physical distancing verification of the area they are filming by offloading the image elaboration to the accelerated edge cloud.

2. Innovation

The demo combines very recent innovations in the communication and service fields. First, the demo exploits huge potentials of accelerated edge cloud. Indeed, accelerated edge cloud can be used not only to run computationally intensive functions at the application level but also computationally intensive 5G functions such as the 5G mobile network stack physical functions. The demo also includes the implementation of a disaggregated Next Generation eNB (gNB) where functions are split into Radio Unit (RU), Distributed Unit (DU), and Central Unit (CU) as currently proposed in the standards [5]. The gNB split Option 2 is considered where the RU includes RF functions only, the DU, which is connected to the RU by an optical fronthaul performs all the functions from the Low-PHY to the Radio Link Control (RLC), and the CU performs all the functions above RLC. Moreover, some of the RAN functions are virtualized and accelerated. In particular, the DU Low-PHY functions are offloaded onto a Field Programmable Gate Array (FPGA) present in the accelerated edge micro data center. Finally, the physical distancing application, based on a slightly modified version of an open source one based on Yolo, exploits a Graphical Processing Unit (GPU) available

in the accelerated edge data center. In summary, the demo shows the benefits of exploiting an accelerated edge cloud for supporting low latency applications.

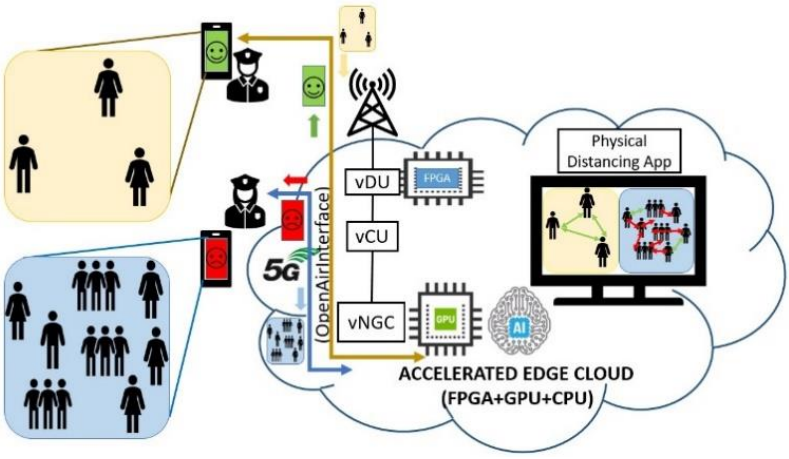

Fig. 1. Demo setup and architecture

3. OFC relevance

The demo presents a multifold relevance to OFC 2021. In general, the demo shows how communications networks can contribute to fighting COVID-19 pandemic.

Moreover, the demo shows that, by exploiting the low latency and broadband connectivity provided by 5G and the computational power offered by accelerated edge clouds, computationally intensive applications can be made available also to end user devices that are not computationally powerful, and battery powered without comprising the application response time.

Furthermore, the demo contributes to shedding some light on the type of vertical applications that can fully benefit by elaboration offloaded onto an accelerated edge cloud.

The demo also shows how the integration of high speed optical fronthaul networks with 5G mobile function virtualization and acceleration can benefit end user applications.

Finally, the demo paves the way to other possible applications supported by an accelerated edge cloud. In the considered demo a 1:1 offloading of the end device applications to the edge cloud is considered. However, the proposed approach can be even more beneficial when images shall be elaborated in a synchronous manner for crowd management. Indeed, in this case, a distributed elaboration by the end devices could be difficult, if not impossible, to achieve. Consider, for example, a situation where different security cameras are monitoring different parts of a smart city. If an accident happens in an area (e.g., a shooting, a fire, etc.), citizens shall be directed how to free the area itself and the nearby areas in the fastest possible ways. In this scenario, a crowd management app deployed in the [i]edge cloud receives the images from the different cameras, elaborates them in real-time by counting the number of people in each area and it correlates the images coming from neighboring areas. Then the app sends notification to the citizen smartphones or to street smart displays or smart traffic signs or smart kiosks to direct them away from the accident area as fast as possible.

4. Implemented software components and Demo description

A single location demo setup is shown in Fig. 2. One or more smartphones are connected to the 5G network running OpenAirInterface (OAI) [6] as mobile software stack. OAI Core Network (CN) is utilized for implementing the NGC functions. Docker Container based virtualized version of DU (vDU), CU (vCU), and NGC (vNGC) components are deployed in the accelerated edge cloud using Dell PowerEdge R740 servers. Ettus X310 Universal Software Radio Peripherals (USRPs) is utilized as gNB radio front-end performing Radio Frequency (RF) functions. In the accelerated edge cloud the vDU Low-PHY functions are offloaded and accelerated onto an FPGA using the OpenCL Framework [7]. The physical distancing app is deployed as well in the accelerated edge cloud, beyond the NGC, and runs in a GPU. An NVIDIA Tesla T4 data center GPU featuring a 320 NVIDIA Turing tensor cores is utilized. NVIDIA T4 supports all AI framework with 50x higher energy efficiency compared to the CPUs due to the low profile PCIe [8].

The physical distancing app is based on an open-source physical distancing app exploiting You Only Look Once (YOLO) [9]. The app is modified to compute the number of social distancing violations and send the notification back to the UE based on a threshold.

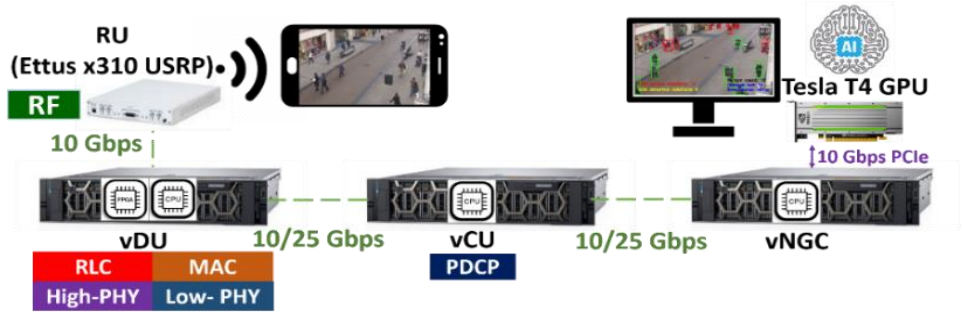

Fig. 2 Physical distancing application on 5G Testbed enabled with FPGA, GPU, CPU and 5G mobile access.

During the demo, images acquired on site are fed through the 5G network for elaboration to the physical distancing app. The results of the elaboration are sent back to the smartphones through the 5G mobile network. The main performance evaluation parameter is the physical distancing app *response time*, computed as the time required to send an image from the smartphone to the physical distancing app and receive the notification to the smartphone. The response time also includes the time required to elaborate the image. For comparison purposes the response time of the demo setup will be compared to the one achieved by performing the elaboration locally in the smartphone. In addition, already available crowd images will be fed together with live images to evaluate the scalability of the proposed solution. It is expected that accelerated edge cloud elaboration can outperform the local image elaboration response time because the higher computation power of the edge cloud can compensate also for the data transfer roundtrip time. However, if the images to be elaborated by the physical distancing app increase (i.e., more smartphones are connected to the same physical distancing app) the response times will get closer. Thus, a scale up of the physical distancing app will be required.

If it will not be possible to deploy the full demo in a single location, alternative deployments will be considered. One is where the scene filming, the RU, and the vDU are deployed in one site while the vCU, vNGC, and application can be deployed at a remote location in the edge cloud. In this latter case the application response time also includes the round-trip time between the two locations. A final alternative, if physical presence will not be possible due to COVID-19 pandemic travel restrictions, a remote audio/video connection to the ARNO testbed in Pisa where the single location demo will be performed will be set up.

Acknowledgements

This study was conducted within the 5G-SOSIA project, which receives funding by the Region of Tuscany under Bando Ricerca Salute 2018, and within the BRAINE project ECSEL JU grant agreement No 876967. The JU receives support from the EU Horizon 2020 research and innovation programme and the Italian Ministry of Education, University, and Research (MIUR).

References

[1] Kashif Bilal, Osman Khalid, Aiman Erbad, Samee U. Khan, “Potentials, trends, and prospects in edge technologies: Fog, cloudlet, mobile edge, and micro data centers”, Computer Networks, Volume 130, 2018, Pages 94-120.
[2] E. Li, L. Zeng, Z. Zhou and X. Chen, Edge AI: On-Demand Accelerating Deep Neural Network Inference via Edge Computing," in IEEE Transactions on Wireless Communications, vol. 19, no. 1, pp. 447-57, Jan. 2020.
[3] N. Dilshad, J. Hwang, J. Song and N. Sung, "Applications and Challenges in Video Surveillance via Drone: A Brief Survey," 2020 International Conference on Information and Communication Technology Convergence (ICTC), Jeju, Korea (South), 2020, pp. 728-732.
[4] K. Bhambani, T. Jain and K. A. Sultanpure, "Real-time Face Mask and Social Distancing Violation Detection System using YOLO," 2020 IEEE Bangalore Humanitarian Technology Conference (B-HTC), Vijiyapur, India, 2020, pp. 1-6.
[5] ITU-T, “Transport network support of IMT-2020/5G,” GSTR-TN5G (2018). Version 1.0.
[6] https://www.openairinterface.org/
[7] F. Civerchia, M. Pelcat, L. Maggiani, K. Kondepu, P. Castoldi and L. Valcarenghi, "Is OpenCL Driven Reconfigurable Hardware Suitable for Virtualising 5G Infrastructure?" in IEEE Transactions on Network and Service Management, vol. 17, no. 2, pp. 849-863, June 2020, doi: 10.1109/TNSM.2020.2964392.
[8] https://www.nvidia.com/content/dam/en-zz/Solutions/Data-Center/tesla-t4/t4-tensor-core-product-brief.pdf
[9] https://github.com/saimj7/Social-Distancing-Detection-in-Real-Time