Enabling Smart Retail through 5G Services and Technologies

Claudio Meani, Pietro Paglierani
R&D
Italtel S.p.A.
Milan, Italy
{claudio.meani, pietro.paglierani}@italtel.com

Athina Ropodi[1], Nikos Stasinopoulos[1], Kostas Tsagkaris[1], Panagiotis Demestichas[1,2]
[1]Incelligent PC, R&D, Athens, Greece
[2]University of Piraeus, Piraeus, Greece
{ar, ns, kt, pd}@incelligent.net

***Abstract*—The advent of 5G has brought great attention on Crowded Events (CE). In this context, this paper presents a novel framework specifically designed for CEs, which combines the contributions of two 5G Application Providers, namely Italtel (a Large Enterprise) and Incelligent (an SME, pursuing its entrance at the scale-up phase). The proposed system can offer to end users a bundle of high-value services, such as high-quality video sharing and personalized recommendations, obtained from historical data and machine learning techniques. In particular, the paper focuses on a smart retail scenario, to generate personalized retail recommendations in a crowded mall or a touristic location. The proposed framework relies on Network Function Virtualization (NFV) and Software Defined Networking (SDN) concepts, and, in line with the Multi-access Edge Computing (MEC) paradigm, leverages compute and storage resources at the network edge. The presented work has been undertaken under the EU Horizon2020 MATILDA project.**

Keywords—NFV, MEC, Machine Learning, High Definition Video Services, Crowded Events, Smart Retail.

I. INTRODUCTION

The advent of 5G networks has brought about a great interest in the high-value services that can be provided during a Crowded Event (CE).

In a CE, a high number of end users concentrate in a small area for a relatively short time, ranging from few hours to a week. Well-known examples of CEs are sport events in stadiums or exhibitions in dedicated venues, but also touristic locations during seasonal peak-periods or malls at peak-hours. During CEs, besides an impressive data traffic growth, one can clearly observe a shift of consumers' behavior, with more video-related activities and social networking, and less voice calls and text messages [1].

However, offering innovative services during CEs poses demanding requirements, which have highlighted the weaknesses of the present telecommunication infrastructure, and have thus played a key role in the definition of the new 5G-architecture [2][3].

5G networks are expected to increase bandwidth and number of connected devices, and decrease latency [3]. To lower equipment costs, increase flexibility and reduce service deployment time, they will strongly rely on SDN and NFV concepts, in a new architectural framework where also MEC will play a major role [4].

This paper describes a novel framework offering high-value, innovative services during a CE. Leveraging NFV and SDN, the solution combines the Italtel i-EVS system for high quality, immersive video services and geo-localization functionalities, with the Incelligent's Artificial Intelligence-based framework which, by analyzing historical and mobility data, can provide personalized recommendations to each single user in real-time. The proposed system has been specifically designed to operate in a smart retail scenario, and can be deployed in a mall or in a touristic location, so that users can benefit of a bundle of innovative services in a straightforward and immediate way.

In line with MEC, the proposed solution exploits compute, storage and networking resources made available by a local virtualized infrastructure, at the network edge [4].

The paper is organized as follows: The next section introduces the smart retail scenario in 5G, and describes the two main components of the system. Section III presents the architectural approach to integrate such components, and discusses benefits, challenges and future directions of the presented Proof of Concept.

II. SMART RETAIL RECOMMENDATIONS SCENARIO IN 5G

A. Scenario workflow

The scenario in play involves a user that is moving around a crowded venue, such as a shopping mall or an open air market area, sharing high quality video with her peers while making purchases around the brick and mortar shops that are part of this venue (Fig. 1).

Fig. 1: A user journey inside a shopping mall area. The user receives high quality media content and personalized recommendations as she moves around the shops

978-1-5386-3873-6/17/$31.00 ©2017 IEEE

The user, through her mobile device, can discover personalized recommendations and offers, based on her exact current location, and her preference to consume certain types of products or services (possibly in sequences or bundles). Such data-driven retail recommendations come in the form of a "Move to next shop that offers an A% discount just for you" visual aid on the user's mobile screen, and are created by applying advanced machine learning methods on the user's purchase history and mobility data, in user-perceived real time. An indicative action and corresponding data flow for the system is depicted in Fig. 2. This representation follows how the network application interprets and reacts to the way a user moves inside the mall and engages in its retail stores.

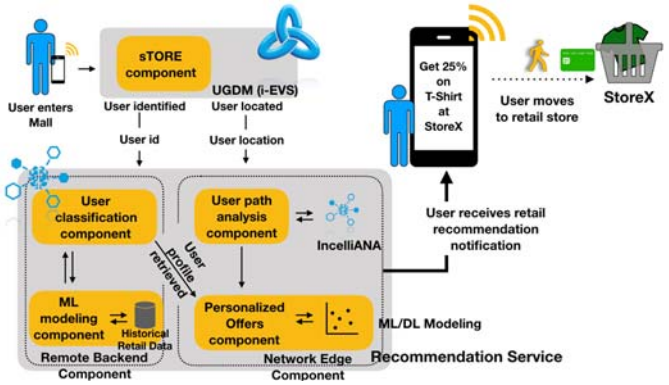

Fig. 2: The work and data flow of the system. A user gets identified, associated with a user profile at the backend component, and receives on screen a personalized/ localized offer by the edge recommendation service component.

B. *Data modeling and ingestion*

In terms of modeling, while there have been different approaches presented in the past [5] -mostly related with traditional statistics and machine learning (ML) methodologies- smart retail recommendations remain a challenge. Additionally, while the Recency-Frequency-Monetary (RFM) modeling has been widely used as an indicative way of describing the client's state and value for the retail industry, it lacks the ability to incorporate relevant client features and the granularity (in terms of time and location) necessary for a smart personalized recommendation mechanism. In the suggested approach, an intelligent recommendation mechanism must include a series of advanced ML methods, incorporating various aspects of the Customer Relationship Management (CRM) modeling. These intelligent methodologies involve data preprocessing, dimensionality reduction, feature selection, advanced customer segmentation and profiling, prediction modeling for customer propensities to buy and near real-time recommendations at the network edge. Interesting to note that the modeling methods take into account information from large datasets including diverse data that involve user specific, spatial and temporal data:

- Client profile (i.e. demographics, registration data, corresponding market segments etc.)
- Time-related data (such as time of year, day, time of day, sales period, mean time spend in specific areas, etc.)
- Area/venue-related data (e.g. mapping of open area/ venue, paths & areas of interest, mapping of stores within the area)
- Current/ previous offers and campaign data (modeled based on parameters such as type of store/offer, monetary value, time of year/ event).

C. *HD video sharing & augmented reality*

1) The i-EVS Framework

i-EVS consists of two components, the i-EVS Virtual Network Function (VNF) and the i-EVS App. The VNF provides video processing and data storage functionalities. It also manages the information related to the users (organized in groups) accessing i-EVS services, by storing and maintaining updated the related data in the i-EVS User and Group Database. The App can be downloaded by the users onto their devices; it offers the possibility to register to i-EVS, and create groups of users who can access the provided services. The App provides geo-localization and image processing services. The main features of the i-EVS VNF and App, as well as their interactions, are summarized in the next paragraphs.

a) The i-EVS VNF

The i-EVS VNF consists of two VNF Components (VNFCs), the Video Transcoding Unit (VTU) and the User and Group Database Manager (UGDM). A simplified block scheme of such VNF is shown in Fig. 3, where also other blocks of the ETSI NFV Management and Orchestration framework are shown. The VTU can provide three main functionalities, i.e. i) Video/ Audio transcoding; ii) Media streaming, and iii) Video/ Audio file upload/download.

The UGDM VNFC, conversely, handles and stores all the information about CE participants, which have registered to i-EVS. It provides Operation and Maintenance (O&M) functionalities, and manages the local storage system where audio/ video data are stored (sTORE), as shown in Fig. 3.

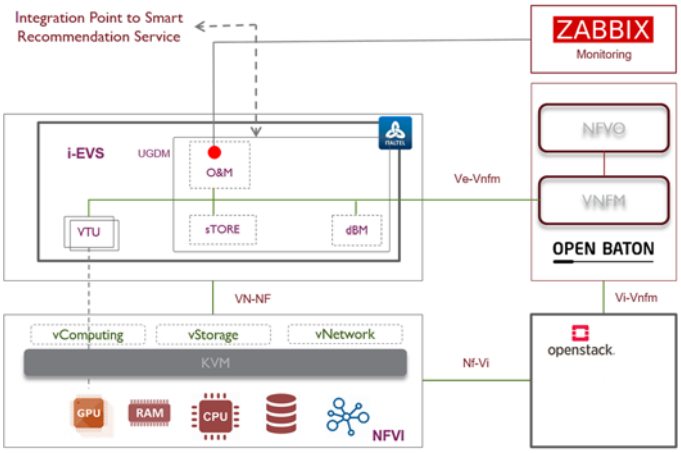

Fig. 3: Block scheme of the i-EVS VNF

b) The VTU VNFC

The VTU can convert audio and video streams from one format to another. The source stream can originate from a file within the local storage system, or as a packetized network stream. The requested transcoding service can be mono-directional, as in video streaming, or bi-directional, like in videoconferencing (see Fig. *4*). In the VTU, the transcoding capabilities are provided by Libav [6], an open source library, which can handle a wide variety of audio/ video coding

standards. For the most compute-intensive video encoding tasks, the VTU relies on Graphical Processing Unit resources [7].

In traditional content delivery systems, users receive the selected contents as video streams. Conversely, i-EVS also gives the possibility to originate and share media contents within a pre-defined group, in real time or as recorded video files in a common storage area. To this end, i-EVS can forward in real-time any received content to any user in a group. Each user receives a notification that a video stream is being originated within her group, and, if interested, can access it in real-time. Originated streams are recorded by the VTU in the distributed storage system, to be shared at a later time. Finally, users with specific rights can send video contents to all the users in the CE. A more detailed description of i-EVS is in [7]. To facilitate sharing of pre-recorded video contents, the VTU provides local storage capabilities. Typically, media contents originated by end users are shared through cloud-based applications, which require the transmission of a file to a centralized cloud infrastructure through the core network.

An interested user can then download that file through the core network. However, in a CE, such operations are usually very slow, due to the high density of users that rapidly saturate the backhaul connection to the core network. Experiments in typical CEs have shown that the upload of a 100MB file at peak hours can take tens of minutes. Conversely, the same upload to the local VTU storage system through the same access network, can take no longer than 10 to 20 seconds.

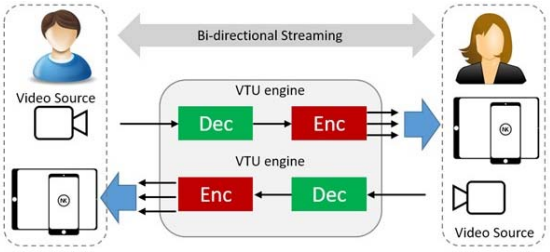

Fig. 4: Simplified Bi-directional VTU Model

c) The UGDM VNF component

UGDM collects and monitors relevant information about CE participants and groups, storing it into an ad hoc database, the User and Group Database. Any user connecting to the network during a CE can download the i-EVS App, and register to the system. Registered users are permanently identified by a randomly generated key that remains unchanged for the entire duration of the event, combined with user name and/or nickname. Through the App and interacting with the UGDM users can create groups. For each user, the database stores the relevant information about identity, group, connectivity and Service Level Agreement (SLA), to guarantee to each user the required type of service. UGDM provides basic security functionalities, such as user password management. In the considered use-case, geo-localization services are provided; the physical position of end users is saved in the database, and continuously updated. In Fig. *3*, UGDM also implements the contact point to interoperate with other functional blocks, such as the Incelligent's Network Edge and Remote Backend Components. In this particular application, it periodically sends specific data for each user, so that personalized recommendations can be originated.

UGDM performs monitoring of the entire system, including VTU performance, through a dedicated block; moreover, it manages the i-EVS storage functionality. Further details on such functional blocks can be found in [7].

2) The i-EVS App

i-EVS offers a web based interface, accessible from any browser. However, to improve User Experience, a specific i-EVS App has been developed.

Through the App, users can register to i-EVS, as described above, create groups and access services. A shared local storage area is assigned to each group, where all the users of the group can read or write media contents privately. The simplest services accessible through the App include video file exchange and video chatting. In addition, geo-location functions and QR code readers can be provided, combined with augmented-reality-based applications.

III. Architectural Approach

A. Short description of Architecture

The core functionalities of the system (described in Fig. *2*) are summarized in the following:

1) User identification

The user must be identified as a single entity across i-EVS and the edge recommendation component. Both components are physically located at the network edge (i.e. at the mall) running as VNF or virtualized containers on an OpenStack edge cloud. The deployment is depicted in Fig. *5*.

A simple scenario always starts with the user registering herself to i-EVS UGDM. The recommendation service component includes a database of previous interactions between consumers and retailers. This database can either be located at the network edge, or at the backend, in the recommendation service providers premises, depending on (mostly privacy conserving) SLAs. This data store contains a combination of historical data, i.e. past purchases made by the considered user.

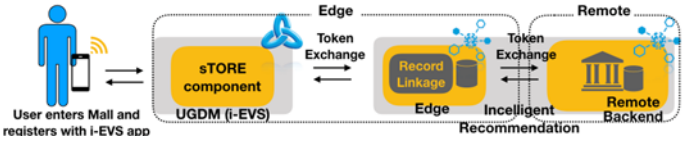

Fig. 5: User identification across the recommendation service edge and backend components

To bridge the data flow between the two services, the i-EVS user securely identifies herself with the edge recommendation component, which acts as a broker to the recommendation service provider. This can happen via a password or fingerprint authentication mechanism on the mobile device. The token returned by a successful identification is stored together with the i-EVS user key, demographics data and location data in the recommendation service database. This data exchange is performed between the two services (i-EVS and recommendation) that live in separate VNF/ containers, and expose interfaces according to the microservice/ service mesh paradigm.

2) User profiling

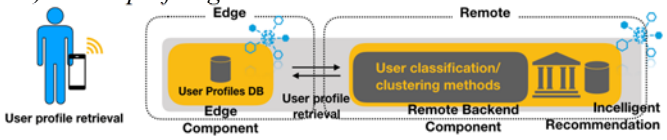

Fig. 6: User profile association via the backend recommendation service component

The user profiling functionality takes place exclusively inside the recommendation service (Incelligent's edge and remote service components) as depicted in Fig. *6*.

Once a user's identity is successfully resolved (as per the previous subsection) her behavior must be modeled. Taking into account the particular demographic data and purchase history, the user is clustered inside a microsegment of the consumer base with specific characteristics and propensities to consume. This modeling task involves advanced machine learning and deep learning methods that (i) are computationally intensive, thus they must be performed where adequate computing power (CPU and GPU based) is available, and (ii) are used as the starting point for a specific offer recommendation, thus they can be run asynchronously with no detrimental effect on the user QoE. Hence, the user profiling functionality can be offloaded to the backend component. The user model, expressed in terms of consumption behavioral patterns, is produced at the remote backend service component; hence, the backend component must expose such a user model to the edge component through a RESTful API with a semi-structured (i.e. JSON) payload. Both service components reside in virtual or container images that are orchestrated externally in terms of availability, service discovery and monitoring.

3) Location-specific retail recommendation feedback

The proposed framework aims at providing retail recommendations to end users through their mobile device. Hence, a user registered to i-EVS services and authenticated by the recommendation service will get offers on her screen.

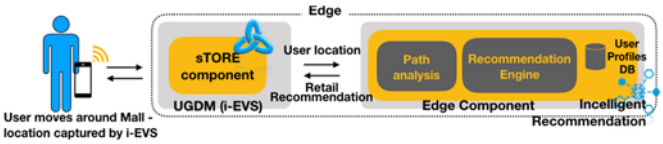

Fig. 7: Retail recommendation associated to the user profile and current location

The visual feedback will be sent by i-EVS in the form of a simple notification on the mobile's notification shade, an overlay on top of the media shared by the i-EVS, or an Augmented Reality layer prompting the user to open his camera and point to relevant brick and mortar shop. This requires interoperability between i-EVS and the recommendation service, analogous to the one described in the User Authentication subsection. To this end, UGDM provides the user id and her most recent location as captured by the i-EVS App running on the mobile device.

On the edge recommendation service component, the user is already registered, and a separate key-value storage is available. In this NoSQL DB (i.e. MongoDB) the key is the user id, while the value is a semi-structured document (i.e. JSON) that contains the modeled user consumer propensities and common paths the user takes. In the next subsection, the offer recommendation functionality is presented. The outcome of that process is returned to i-EVS as a key-document pair (in JSON) of small payload with the key being the user id and value document a set of best retail recommendations.

a) Computing the final recommendations

At the edge, the location-specific recommendation is calculated given the exact user location and the modeled propensities. For this, a separate compute instance is deployed inside an orchestrated container. Its role is first and foremost to return the personalized recommendation to the i-EVS request and, secondarily, to return the same prediction to the NoSQL DB to retrain the model located at the backend at a later time and to keep the prediction result inside an in-memory DB (i.e. Redis) for fast access.

Computing the final recommendation is essentially a stateless function and can thus be readily scaled in/ out in an orchestrated deployment. Monitoring service and network Key Performance Indicators such as latency and number of concurrent requests will drive the scaling process managed by the system orchestrator.

B. Infrastructural requirements

The descriptions given in the previous sections allow to summarize the main infrastructural elements needed to deploy the proposed system.

1) Compute and Storage Resources

The i-EVS VNF and the recommendation service edge component run as software appliances on the edge virtualized infrastructure, which must hence provide adequate local compute and storage resources, as envisioned by MEC. For video processing functions, GPUs must be available [7]. Such GPU resources can also be used by the edge recommendation service component. The virtualized infrastructure manager, as shown in Fig. *3*, is OpenStack. The system can interoperate with different ETSI MANO Open Source orchestrators, such as OpenBaton or Open Source Mano. Additionally, remote resources (i.e. not included in the local MEC infrastructure), to store media contents produced by the users also beyond the CE duration and to run the backend recommendation service, can be eventually required.

2) Connectivity

Wireless local connectivity must be provided to the users to access the provided services. Moreover, connectivity of the local blocks to the remote storage and compute resources are needed. To provide high quality, immersive video services, anywhere and with any device, two features play a fundamental role from the local connectivity point of view, i.e., bandwidth, and latency, both on the user and on the control plane.

The 5G architecture will provide significant advances related to such parameters. 5G networks will enable eMBB (enhanced Mobile Broad Band) types of services, and will operate in scenarios with high user device densities (more than 10000 per km^2), and low latency [2].

The 5G network will start operations from 2020. Thus, other scenarios must be considered. One solution consists in using a WIFI access network. A second option is the use of Small Cells.

In this case, the present 4G mobile network architecture can be used. For instance, a specific solution for a 4G scenario can be found in [8]. Standardization aspects related to the use of i-EVS in 4G and 5G scenario are discussed in [9].

C. Descriptive analytics

The modeling of the recommendation process may vary according to the actual limitations on hardware, software and the modeling approach itself with respect to descriptive, predictive and prescriptive analytics. In fact, the complexity of modeling such high-volume, as well as diverse and complex data, as described in section II.B, requires hardware resources that are not necessarily available at the edge. At the same time, the data employed ideally include information from multiple open areas/ markets and venues (e.g. malls, stadiums) for the same recommendation service provider and therefore putting an unnecessary strain on network resources for data exchange.

For this reason, the majority of the modeling tasks is performed remotely in the retailer's central premises and only the venue specific models are deployed at the network edge (path analytics/ mobility and location-time-offer-dependent modeling). In particular, models are trained at the central premises using open source and proprietary software for both supervised and unsupervised techniques. Incelligent proprietary software enables automatic selection the best approach among various methodologies. Various data analysis methodologies are employed and tested -supervised and unsupervised, separately or combined- for clustering, dimensionality reduction, classification and regression, such as K-means, Principal Component Analysis, t-Distributed Stochastic Neighbor Embedding, Gaussian Mixtures Models, Neural Networks, Linear/ Logistic Regression, Decision Trees, Gradient Boosting Trees etc.[10]-[13]. At the headquarters deep neural networks [13], and specifically Long Short Term Memory (LSTM) Recurrent Neural Networks for sequence prediction for regression and classification with time series data are employed to calculate user propensities to buy at specific times. These models are trained at regular intervals and offer real-time propensity model results.

On the other hand, at the edge, basic user profile data, segmentation and location analytics for the extraction of highly probable paths and mobility patterns along with the area map representation of the shop placement and type along with the corresponding offers are employed for user movement prediction. These mobility patterns can be regularly updated based on new mobility patterns or with respect to specific events (e.g. concerts). Approaches ranging from rule-based methods to deep reinforcement learning (DRL) are considered to deliver the retail recommendation offer. At this stage, the predictions are formulated as a case of a Markov Decision Process and/or a sequence classification (LSTM) regarding the probability and size of transaction given specific nearby offers and the advanced user profile assigned by the headquarters. However, a possible extension of this approach is the incorporation of a DRL framework from observed samples based on user's current state (profile, location, etc.) and a discrete action space, as proposed in a similar study [14].

IV. Conclusion

The paper has presented a framework designed to provide high value, innovative services to end users during crowded events. The system combines the i-EVS framework by Italtel, for immersive video services, and the Incelligent system, which can provide personalized recommendations to end users in real time from the analysis of historical and mobility data using machine learning techniques. The architectural approach to combine the two systems has been described and discussed. The integration process relies on NFV and SDN concepts, and, in line with MEC, exploits virtualized compute, storage and networking resources at the network edge. The presented services can be provided with WIFI and 4G small cell technology, but will demonstrate their full potentialities in a 5G environment.

Acknowledgment

This work has been undertaken under the EU Horizon2020 MATILDA Project (Grant Agreement 761898).

References

[1] Ericcson, "Rio-aiming-higher-report", 2016 https://www.ericsson.com/assets/local/narratives/networks/documents/rio-aiming-higher-report.pdf

[2] European Commission, 5G Infrastructure Public Private Partnership (PPP): The next generation of communication networks will be Made in EU. Digital agenda for Europe. Technical Report, Feb. 2014.

[3] NGMN: 5G White pape, 2015,. [Online]. Available at: https://www.ngmn.org/fileadmin/ngmn/content/downloads/Technical/2015/NGMN_5G_White_Paper_V1_0.pdf

[4] ETSI: Mobile-Edge Computing - Introductory Technical White Paper (2014).

[5] K. Tsiptsis & A. Chorianopoulos, "Data Mining in CRM". In "Data Mining Techniques in CRM", John Wiley & Sons, Ltd., 2010, pp. 1–15.

[6] Libav. [Online]. Available at: http://libav.org/documentation/

[7] A. Albanese, P. S. Crosta, C. Meani, & P. Paglierani, "Immersive Video Services at the Edge: an Energy-Aware Approach", IARIA International Journal on Advances in Telecommunications, vol 10, n. 3&4, January 2017

[8] Fajardo et al., "Introducing Mobile Edge Computing Capabilities through Distributed 5G Cloud Enabled Small Cells", Mobile Netw Appl, vol. 21(4), pp. 564–574, Aug. 2016

[9] B. Blanco et al., "Technology pillars in the architecture of future 5G mobile networks: NFV, MEC and SDN," Computer Standards & Interfaces, Available online 4 January 2017, ISSN 0920-5

[10] T. Hastie, R. Tibshirani, & J. Friedman, "The Elements of Statistical Learning" (2nd ed.), Springer New York, 2009.

[11] L.v.d. Maaten & G. Hinton, "Visualizing data using t-sne" Journal of Machine Learning Research, Vol 9, pp. 2579—2605, Nov. 2008

[12] C. Bailey, P.R. Baines, H. Wilson & M. Clark, "Segmentation and customer insight in contemporary services marketing practice: why grouping customers is no longer enough". Journal of Marketing Management, vol. 25 (3-4), pp.2 27–252, 2009

[13] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. "Gradient-based learning applied to document recognition." *Proceedings of the IEEE, vol.86(11, pp. 2278-2324*, November 1998

[14] Y. Tkachenko. "Autonomous CRM Control via CLV Approximation with Deep Reinforcement Learning in Discrete and Continuous Action Space". CoRR, abs/1504.01840, 2015.