Programmable Edge-to-Cloud Virtualization for 5G Media Industry: The 5G-MEDIA Approach

Stamatia Rizou[1(✉)], Panagiotis Athanasoulis[1], Pasquale Andriani[2], Francesco Iadanza[2], Panagiotis Trakadas[3], David Griffin[4], Morteza Kheirkhah[4], David Breitgand[5], Avi Weit[5], Refik Fatih Ustok[6], Selcuk Keskin[6], Francesca Moscatelli[7], Giacomo Bernini[7], Gordana Macher[8], Javier Serrano[9], and David Jimenez[9]

[1] Singular Logic S.A., Athens, Greece
srizou@singularlogic.eu, pathanasoulis@ep.singularlogic.eu
[2] Engineering, Rome, Italy
{pasquale.andriani,francesco.iadanza}@eng.it
[3] National and Kapodistrian University of Athens, Athens, Greece
ptrakadas@uoa.gr
[4] University College London, London, UK
{d.griffin,m.kheirkhah}@ucl.ac.uk
[5] IBM Israel Ltd Haifa, Haifa, Israel
{davidbr,weit}@il.ibm.com
[6] Netas, Istanbul, Turkey
{fustok,selcukk}@netas.com.tr
[7] Nextworks, Pisa, Italy
{f.moscatelli,g.bernini}@nextworks.it
[8] Institut für Rundfunktechnik, Munich, Germany
gordana.macher@irt.de
[9] Universidad Politecnica de Madrid, Madrid, Spain
{jsr,djb}@gatv.ssr.upm.es

Abstract. To ensure high Quality of Experience (QoE) for end users, many media applications require significant quantities of computing and network resources, making their realization challenging in resource constrained environments. In this paper, we present the approach of the 5G-MEDIA project, providing an integrated programmable service platform for the development, design and operations of media applications in 5G networks, facilitating media service management across the service life cycle. The platform offers tools to service developers for efficient development, testing and continuous correction of services. One step further, it provides a service virtualization platform offering horizontal services, such as a Media Service Catalogue and accounting services, as well as optimization mechanisms to flexibly adapt service operations to dynamic conditions with efficient use of infrastructure resources. The paper outlines three use cases where the platform was tested and validated.

Published by Springer Nature Switzerland AG 2020
I. Maglogiannis et al. (Eds.): AIAI 2020 Workshops, IFIP AICT 585, pp. 95–104, 2020.
https://doi.org/10.1007/978-3-030-49190-1_9

Keywords: Network function virtualization · Service Virtualization Platform · Cognitive network optimizer · Quality of experience

1 Introduction

Media applications are among the most demanding applications in terms of QoS, since they require the near real-time delivery and processing of media content across dispersed geographical locations. Recent advancements in 5G technologies [19,23,24] can provide valuable solutions for the realization of complex demanding media-related scenarios enabling the timely media content delivery and processing though intelligent and flexible service orchestration [15,20,22].

The 5G PPP 5G-MEDIA Phase 2 project has worked in the design and implementation of a platform that supports the media services lifecycle management providing a holistic solution encompassing mechanisms and tools for the development, testing and continuously-optimized deployment of media services. The main innovations of the proposed platform lie in the following parts: i) the offering of a complete Service Development Kit (SDK), encompassing tools for the media service validation and emulation as well as for the testing of optimization algorithms in emulation environment and the continuous corrective sizing of resources required after the media service deployment; ii) the development of two auxiliary services, the 5G-MEDIA Service Catalogue and the AAA mechanisms facilitating the media service management as well as accounting and billing mechanisms; iii) the introduction of a multi-hierarchical cognitive network optimizer catering for the continuous optimization of media services during and after deployment and iv) the enablement of traditional as well as serverless orchestration allowing event-driven orchestration of services at run time. While the main innovation aspects of the 5G-MEDIA solution can be applied to other application domains beyond the media industry, the 5G-MEDIA project has focused on the validation and evaluation of its solution in the media vertical, providing promising results for the value and use of the platform.

In this paper, we present the 5G-MEDIA technical solution and we discuss the functionalities of its technical components. We provide a brief overview of the main benefits achieved in the context of the use cases addressed in the project, although a deeper analysis of the results of the validation and trials testing of the solution is beyond the scope of this paper. Section 2 provides an overview of the refined high level 5G-MEDIA architecture and explains its main components comprised by the 5G-MEDIA Service Development Kit (SDK) and the Service Virtualization Platform (SVP). Section 3 explains the benefits achieved during pilot testing, before concluding the paper in Sect. 4.

2 5G-MEDIA High Level Architecture

In this Section, the main components of the 5G-MEDIA architecture shown in Fig. 1 are presented.

2.1 Service Development Kit

The SDK can be considered as a sandbox where the service developers can develop, emulate and test their NFV-based services, integrating both development and operational aspects together.

The main tools offered to the service developers within the 5G-MEDIA SDK are as follows: i) **Packaging Tools:** These collect the component software artefacts (e.g. required libraries) and generate software packages to be uploaded to the Private Catalogue. A range of packaging options are supported to equip the SDK to harmonize various VNF types such traditional plain ISO-based, unikernel-based and container-based VNFs; ii) **Validator:** This tool is the main interface for the developers to write/edit both NSDs and VNFDs, and validate them against TOSCA-based [8] and/or OSM-IM based [5] schemas. Validated TOSCA-based descriptors are then onboarded to the private catalogue via the validator web user interface (UI); iii) **Editor:** The editor is a web-based application whose main purpose is to assist with creating and editing of applications and their descriptors; iv) **Emulator:** The emulator in 5G-MEDIA SDK facilitates local prototyping and testing of real network functions in emulated network

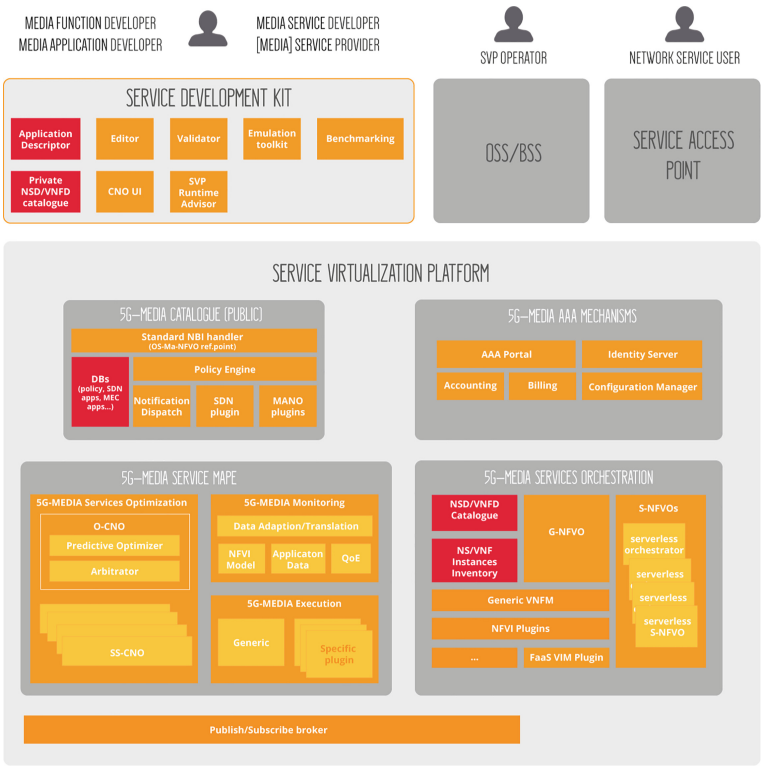

Fig. 1. High-level 5G-MEDIA architecture

topologies running on developers' machines. It leverages the vim-emu for non-FaaS VNFs [14] and the FaaS-vim for the FaaS VNFs [14]; **Benchmarking:** The 5G-MEDIA SDK benchmarking tool supports load-testing under several resource constraints on VNFs that are deployed on the emulation environment. Application developers can benefit from these tests for finding bugs, detecting congestions or investigating issues in their applications; **CNO Training:** Using this tool, the developer can configure a training model based on reinforcement learning, view the training performance and if the model achieves desirable performance, the developer can deploy it to SVP in order to have it utilized by the CNO; **SVP Runtime Advisor:** The SVP Runtime Advisor aims to feed the developer with necessary information about how their services are working in the operational environment. This is particularly necessary when the emulation environment capabilities are significantly different from those in the operational environment.

2.2 5G-MEDIA Service Virtualization Platform

This subsection provides further details on the design and implementation for the core components of the SVP [16].

5G-MEDIA Public Catalogue. The 5G Apps and Services Catalogue is a structured repository of heterogeneous media applications, functions and service descriptors based on a generic and unified standard format, such as VNF Descriptor (VNFD) and NS Descriptor (NSD) data models based on ETSI GS NFV SOL001 v2.5.1 [3] and VNF Package structure based on ETSI GS NFV SOL004 v2.5.1 [4]. The catalogue offers NFV MANO and domain-specific translation functions from ETSI NFV standard to specific or proprietary descriptors and packages format. In the context of 5G-MEDIA the 5G Apps and Services Catalogue provides an automated translation and mapping from the ETSI NFV standard TOSCA based VNF Descriptor and NS Descriptors to the ETSI OSM [5] specific data models, being OSM the reference NFV MANO tool in the project.

Besides the above features, the 5G App and Service Catalogue also connects the DevOps and production environments. To this end a private 5G App and Service Catalogue instance running in the local developer environment within the 5G-MEDIA SDK is connected through a dedicated catalogue-to-catalogue plugin to the production (public) 5G App and Service Catalogue instance running in the 5G-MEDIA SVP with the aim of publishing the designed validates media services and VNFs in the production platform.

5G-MEDIA AAA Services. The main role of the AAA portal is the provisioning of a centralised and simplified view to administrators of resource utilisation and the associated costs for services running on the platform. It has been designed to allow future extensibility through a plug-in architecture. It currently supports OpenID Connect protocol to facilitate the integration of additional

resources and the two login profiles [12] for UI and API access, "Authorisation Code Flow" and "Resource Owner Password Grant Code Flow". The inner AAA architecture comprehends an Identity Server based on KeyCloak [9] that provides the access and refresh tokens to the authorised users. All the configuration is managed by the AAA portal that takes in charge the setup on the different resources involved: Identity Server, NFVO/VNFM, VIM, Catalogue, SDK, MAPE.

5G-MEDIA Service MAPE. The Media Service Monitoring, Analysis, Planning and Execution (MAPE) components provide the intelligence to dynamically manage and provide infrastructure resources for the deployed media services according to observed changes in user demand patterns, availability and performance of network and computational resources.

QoE Monitoring. Besides the collection of resource consumption data, the 5G-MEDIA project has defined application-specific metrics, relevant to media use cases for quantifying the Quality of Experience (QoE). In order to measure the performance of the service in terms of perceived quality, the platform deploys a QoE probe as a VNF close to the end user, which provides a Mean Opinion Score (MOS) through a statistical model fed by Non-Referenced (NR) video metrics. The goal is achieved by using different ML algorithms in order to obtain the estimated QoE from different data coming from the parameterisation of video services such as bitrate, intrinsic video features as temporal and spatial complexity, the aforementioned NR metrics as blurring. Finally, video quality losses are obtained by applying specific indicators proposed by AGH University [13]. The resulting QoE data are considered in the MAPE optimization to maximize perceived quality at application level.

Cognitive Network Optimizer. For the analysis of monitoring data, 5G-MEDIA adopts a hierarchical Cognitive Optimizer model to allocate and distribute resources across its services. Without loss of generality, we categorize this hierarchy into two broad levels and we name them as follows: (1) Overarching-CNO (O-CNO); and (2) Service Specific-CNO (SS-CNO). In a nutshell, the former optimises resources across all services while the latter optimises the allocated resources to a service (granted by the O-CNO). O-CNO conducts two separate but complementary tasks, each of which may follow separate optimisation algorithm. First task is an offline task related to planning, predicting and allocating resources to *all* services periodically (potentially with a long interval, e.g., in the order of hours or days). In 5G-MEDIA this has been implemented with a genetic algorithm which tries to optimally allocate resources to services based on their predicted resource consumption/demand [21]. We refer to this component of O-CNO as the predictive optimizer. The second task allocates resources on demand from a pool of resources shared across all services (or at least shared across a subset of all services). This is implemented as a utility based optimization algorithm which also considers service priorities, constraints and service level agreements.

The latter algorithm is mainly responsible to dynamically allocate shared-pool resources between individual services under the control of the lower-level SS-CNO. Unlike the predictive optimizer side of O-CNO, this is a run-time online algorithm that reacts to requests in real-time.

5G-MEDIA Service Orchestration. In 5G-MEDIA, we adopt an approach that combines traditional VNF orchestration with a novel approach that we term *Serverless Orchestration*. This resulted in an integrated 5G-MEDIA Service Orchestration platform that uses existing VIMs supporting OpenStack and implementing new VIM to support OpenNebula integration. Serverless VNFs are characterised by being invoked in response to service-specific events rather than being instantiated as part of the traditional network service instantiation flows. There are no provisions in the current ETSI standards for orchestrating serverless VNFs. To handle these complex run-time orchestration flows the project has developed a novel event-driven serverless orchestrator (based on Argo Workflows [2] and Argo Events [1] projects) – resulting in a full serverless management and orchestration stack in 5G-MEDIA Service Virtualization Platform. This stack is interoperable with OSM and complements its functionality allowing to unleash the full potential of a serverless computing paradigm for media applications.

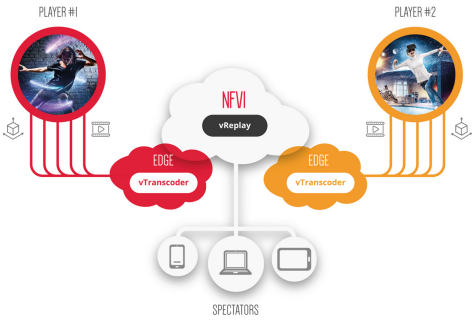

Fig. 2. Tele-immersive gaming scenario

3 5G-MEDIA Use Cases

In the following, we outline the scope of the 5G-MEDIA use cases [17] and the main benefits achieved.

3.1 Tele-Immersive Gaming

This use case demonstrates a real-time interactive immersive media application [18] in which the two players are volumetrically reconstructed and interact with each other in a common virtual gaming environment through their digitized virtual representations (i.e. textured 3D shapes). The application allows for the

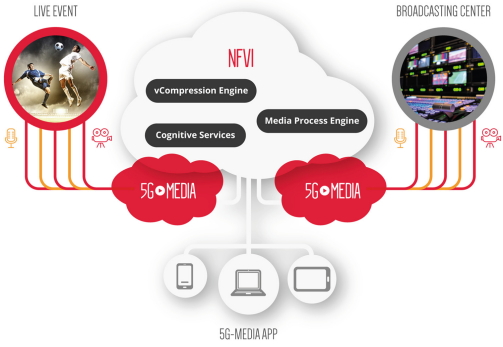

Fig. 3. Remote production scenario

live spectating of each gaming session by remote third party users. The high-level scenario is illustrated in Fig. 2. The role of the 5G-MEDIA platform in this use case lies in the flexible real-time adaptive streaming service, through the use of serverless orchestration that enables the on-demand deployment of transcoding services upon the start of a game session or for the delivery of replay clips to spectators. This use case has tested the efficacy of the CNO considering two different scenarios: i) the optimized selection of transcoding profile to ensure QoE satisfaction; ii) the optimization of the cost efficiency of each session in order to eventually maximize profits, while also retaining the serviced QoE levels relying on the interactions between O-CNO and SS-CNO. The latter approach demonstrated the adoption of a finer-grained service cost analysis opening up new opportunities for application centric optimization.

3.2 Mobile Contribution, Remote and Smart Media Production

This use case enables two media production scenarios. Scenario 1) is a remote production of an event without the need for dedicated infrastructure to be specifically deployed in the event venue as shown in Fig. 3. Scenario 2) considers the streaming of live events via smartphones or tablets by spectators or journalists for enhancing the program stream of an event. The role of the 5G-MEDIA Platform in this use case is to ensure that the media processing functions are efficiently deployed in cloud infrastructure enabling low latency and high throughput as required by live streaming and media processing. This is achieved by using the 5G-MEDIA Service MAPE to optimize bitrate/compression levels of media streams and ensure QoE for the realization of the remote production scenario. In addition, this use case demonstrates the use of FaaS orchestration by automatically deploying media services upon the start of a mobile contribution session. Serverless VNFs implementing cognitive services (captions production and face recognition) for mobile contribution as well as video stream vSplitter and sink are being deployed on demand at the edge where and when they are needed.

During the project the "Remote Production" scenario was approved by supporting a real production of RTVE's Radio3. A time difference of around t = 500 ms, i.e. around 12 frames, was measured in the pretest and could be decreased to 10 frames during the final remote production. The work of the CNO was approved in a "Multi-Instance" scenario, where UC1 and the "Mobile Contribution" scenario shared resources (e.g. GPUs) in the same Edge. In this scenario, the resources where allocated by the CNO.

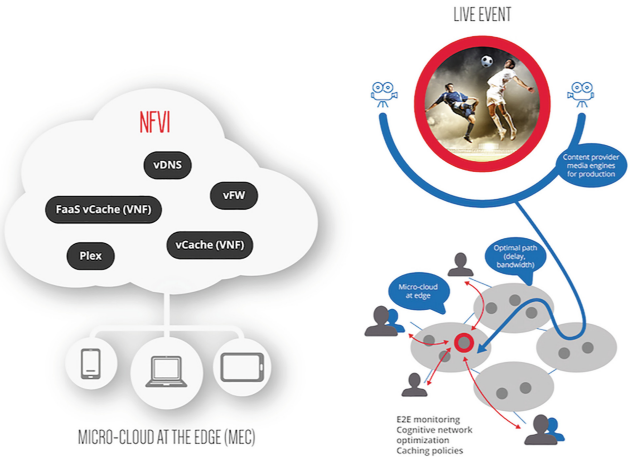

Fig. 4. UHD over vCDN scenario

3.3 Ultra HD over Content Delivery Networks

The focus of this scenario is enabling Media Service Providers (MSPs) to build flexible and adaptable media distribution service chains, made up of virtualized functions, and deliver UHD media contents while users are moving in a geographically distributed 5G network, as shown in Fig. 4. The 5G-MEDIA SDK and Services Catalogue are used for the design, dimensioning, creation and onboarding of the media services, such as the media caching VNFs building the vCDN service. In practice, this service is realized as a multi-NFVI scenario, where regular VM-based media caches (running in Openstack [10]) are integrated in the same media service instance with FaaS based media caches (running in Kubernetes [6]) to create a hierarchical media delivery chain. At the root, a media origin server (based on Plex [11]) hosts all of the media contents that can be accessed by end-users from their fixed (e.g. using a Nextworks Symphony entertainment platform audio/video client [7]) or mobile devices. The 5G-MEDIA SVP is used for the instantiation and application configuration of UHD media delivery services as NFV Network Services. Here, the coordinated configuration of the media caches and load balancing functions is applied by the 5G-MEDIA SVP to re-direct users to proper media content delivering caches (e.g. based on load). Moreover, the 5G-MEDIA MAPE

module monitors the running services and predicts through online ML algorithms operation anomalies (at both network and media service level) due to service congestion (e.g. caused by flash crowds). Upon such predictions, optimization events are generated to dynamically scale the media caches hierarchy (with the option of creating either regular VM-based or serverless virtualized caches) and properly reconfigure the load balancing function for maintaining user perceived quality.

3.4 Cross-Cutting Scenarios

To validate the efficacy of the 5G-MEDIA platform, the project has considered two cross-cutting scenarios mixing more than one sessions or service in the same infrastructure. In particular, to showcase the behaviour of 5G-MEDIA Service MAPE in a multi-tenant environment, the use of O-CNO and SS-CNO was tested in the presence of multiple sessions of the same use case (e.g., parallel remote production sessions sharing the same edge and the same infrastructure) or different use cases (e.g., co-existence of mobile contribution use case and Tele-immersive use case) competing for limited resources such as GPUs in the same NFVI. These test scenarios show the applicability of the proposed approach and in particular of the CNO architecture in multi-tenant environments.

4 Conclusions

This paper presents the 5G-MEDIA high level architecture presenting the role and functionalities of its components. Overall, the 5G-MEDIA approach supports the management of media services across their life cycle, by providing a Service Development Kit for the development and validation of services and a Service Virtualization platform for their continuous optimized deployment. The paper also provides a brief description of the media use cases and the role of the 5G-MEDIA platform in optimizing their performance.

References

1. Argo Events. https://argoproj.github.io/projects/argo-events
2. Argo Workflows. https://argoproj.github.io/argo
3. ETSI GS NFV-SOL 001 V2.5.1: Protocols and data models; NFV descriptors based on Tosca specification. ETSI NFV ISG
4. ETSI GS NFV-SOL 004 V2.5.1. Protocols and data models; VNF package specification. ETSI NFV ISG
5. ETSI Open Source Mano. https://osm.etsi.org/
6. Kubernetes. https://kubernetes.io
7. Nextworks Symphony Entertainment Platform. http://www.nextworks.it/en/products/brands/symphony
8. Oasis Tosca: Tosca simple profile for network functions virtualization. http://docs.oasis-open.org/tosca/TOSCA-Simple-Profile-YAML/v1.0/TOSCA-Simple-Profile-YAML-v1.0.pdf. Accessed 2 Mar 2020

9. Open Source Identity and Access Management. https://www.keycloak.org. Accessed 2 Mar 2020
10. Openstack. https://www.openstack.org
11. Plex Media Server. https://www.plex.tv
12. The OAuth 2.0 Authorization Framework Section 1.3. https://tools.ietf.org/html/rfc6749#section-1.3. Accessed 2 Mar 2020
13. Video Quality Indicators - AGH University. http://vq.kt.agh.edu.pl/metrics.html
14. Acar, U., Ustok, R., Keskin, S., Breitgand, D., Weit, A.: Programming tools for rapid NFV-based media application development in 5G networks. In: IEEE Conference on Network Function Virtualization and Software Defined Networks (NFV-SDN), Verona, Italy, pp. 1–5 (2018)
15. Alemany, P., et al.: Network slicing over a packet/optical network for vertical applications applied to multimedia real-time communications. In: 2019 IEEE Conference on Network Function Virtualization and Software Defined Networks (NFV-SDN), pp. 1–2. IEEE (2019)
16. Alvarez, F., et al.: An edge-to-cloud virtualized multimedia service platform for 5G networks. IEEE Trans. Broadcast. **65**(2), 369–380 (2019)
17. Caruso, G., et al.: Embedding 5G solutions enabling new business scenarios in media and entertainment industry. In: 2019 IEEE 2nd 5G World Forum, pp. 460–464. IEEE (2019)
18. Karakottas, A., Papachristou, A., Doumanoqlou, A., Zioulis, N., Zarpalas, D., Daras, P.: Augmented VR. In: 2018 IEEE Conference on Virtual Reality and 3D User Interfaces (VR), p. 1. IEEE (2018)
19. Peuster, M., et al.: Introducing automated verification and validation for virtualized network functions and services. IEEE Commun. Mag. **57**(5), 96–102 (2019)
20. Pol, A., et al.: Advanced NFV features applied to multimedia real-time communications use case. In: 2019 IEEE 2nd 5G World Forum (5GWF), pp. 323–328. IEEE (2019)
21. Rocha, M., Phan, T.K., Reis, J., Griffin, D., Rio, M.: Triptych: multi-objective optimisation of service deployment costs, application delay and bandwidth usage. In: 2019 IFIP Networking Conference (IFIP Networking), pp. 1–9 (2019)
22. Shekhawat, Y., et al.: Orchestrating live immersive media services over cloud native edge infrastructure. In: 2019 IEEE 2nd 5G World Forum (5GWF), pp. 316–322. IEEE (2019)
23. Soenen, T., et al.: Insights from Sonata: implementing and integrating a microservice-based NFV service platform with a DevOps methodology. In: 2018 IEEE/IFIP Network Operations and Management Symposium, NOMS 2018, pp. 1–6. IEEE (2018)
24. Trakadas, P., et al.: Comparison of management and orchestration solutions for the 5G era. J. Sens. Actuator Netw. MDPI **9**(1), 1–4 (2020)