An Edge-to-Cloud Virtualized Multimedia Service Platform for 5G Networks

Federico Alvarez, David Breitgand, David Griffin, Pasquale Andriani, Stamatia Rizou, Nikolaos Zioulis, Francesca Moscatelli, Javier Serrano, Madeleine Keltsch, Panagiotis Trakadas, T. Khoa Phan, Avi Weit, Ugur Acar, Oscar Prieto, Francesco Iadanza, Gino Carrozzo, Harilaos Koumaras, Dimitrios Zarpalas, and David Jimenez

***Abstract*—The focus of research into 5G networks to date has been largely on the required advances in network architectures, technologies, and infrastructures. Less effort has been put on the applications and services that will make use of and exploit the flexibility of 5G networks built upon the concept of software-defined networking (SDN) and network function virtualization (NFV). Media-based applications are amongst the most demanding services, requiring large bandwidths for high audio–visual quality, low-latency for interactivity, and sufficient infrastructure resources to deliver the computational power for running the media applications in the networked cloud. This paper presents a novel service virtualization platform (SVP), called 5G-MEDIA SVP, which leverages the principles of NFV and SDN to facilitate the development, deployment, and operation of media services on 5G networks. The platform offers an advanced cognitive management environment for the provisioning of network services (NSs) and media-related applications, which directly link their lifecycle management with user experience as well as optimization of infrastructure resource utilization. Another innovation of 5G-MEDIA SVP is the integration of serverless computing with media intensive applications in 5G networks, increasing cost effectiveness of operation and simplifying development and deployment time. The proposed SVP is being validated against three media use cases: 1) immersive virtual reality 3-D gaming application; 2) remote production of broadcast content incorporating user generated contents; and 3) dynamically adaptive content distribution networks for the intelligent distribution of ultrahigh definition content. The preliminary results of the 5G-MEDIA SVP platform evaluation are compared against current practice and show that the proposed platform provides enhanced functionality for the operators and infrastructure owners, while ensuring better NS performance to service providers and end users.**

***Index Terms*—5G networks, network functions virtualization, serverless computing, immersive media, remote production, content delivery networks.**

Manuscript received October 27, 2018; revised December 21, 2018; accepted January 30, 2019. This work was supported in part by the European Commission H2020 Programme, 5G-MEDIA Project under Grant 761699. Parts of this paper have been published in the Proceedings of the IEEE BMSB 2018, Valencia, Spain. *(Corresponding author: Federico Alvarez.)*

F. Alvarez is with the GATV Research Group, Universidad Politécnica de Madrid, 28040 Madrid, Spain (e-mail: fag@gatv.ssr.upm.es).

D. Breitgand and A. Weit are with Computing as a Service, IBM Haifa Research Labs, Haifa 3498825, Israel (e-mail: davidbr@il.ibm.com; weit@il.ibm.com).

D. Griffin and T. K. Phan are with the Institute of Communications and Connected Systems, Department of Electronic and Electrical Engineering, University College London, London WC1E 7JE, U.K. (e-mail: d.griffin@ucl.ac.uk; t.phan@ucl.ac.uk).

P. Andriani and F. Iadanza are with the Research and Development Laboratory, Engineering Ingegneria Informatica S.p.A., 00144 Rome, Italy (e-mail: pasquale.adriani@eng.it; francesco.iadanza@eng.it).

S. Rizou is with the European Projects Department, Singular Logic, 145 64 Athens, Greece (e-mail: srizou@singularlogic.eu).

N. Zioulis is with the Information Technologies Institute, Centre for Research and Technology Hellas, 57001 Thessaloniki, Greece, and also with the Signals, Systems and Radiocommunications Department, Universidad Politécnica de Madrid, 28040 Madrid, Spain (e-mail: nzioulis@iti.gr).

F. Moscatelli and G. Carrozzo are with NextWorks, 56122 Pisa, Italy (e-mail: f.moscatelli@nextworks.it; g.carrozzo@nextworks.it).

J. Serrano and D. Jimenez are with the Signals, Systems and Radiocommunications Department, Universidad Politécnica de Madrid, 28040 Madrid, Spain (e-mail: jsr@gatv.ssr.upm.es; djb@gatv.ssr.upm.es).

M. Keltsch is with the Research and Development Department, Institut für Rundfunktechnik, 80939 Munich, Germany (e-mail: keltsch@irt.de).

P. Trakadas is with the Department of Electrical Engineering, Technological and Educational Institute of Sterea Ellada, 34100 Psachna, Greece (e-mail: ptrakadas@teiste.gr).

U. Acar is with the Multimedia Research and Development Department, NETAS, Istanbul 34912, Turkey (e-mail: uacar@netas.com.tr).

O. Prieto is with the Network Department, Radio Televisión Española, 28007 Madrid, Spain, and also with the Signals, Systems and Radiocommunications Department, Universidad Politécnica de Madrid, 28040 Madrid, Spain (e-mail: oscar.prieto@rtve.es).

H. Koumaras is with the Institute of Informatics and Telecommunications, NCSR Demokritos, 15310 Athens, Greece (e-mail: koumaras@iit.demokritos.gr).

D. Zarpalas is with the Information Technologies Institute, Centre for Research and Technology Hellas, 57001 Thessaloniki, Greece (e-mail: zarpalas@iti.gr).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TBC.2019.2901400

I. Introduction

A GREAT deal of work is currently underway to organize 5G technologies in supporting ultra-high-quality streaming media and entertainment applications. In this direction, virtualization and flexible scaling of cloud resources and Network Services (NSs), at both the network core and edge, will be the key elements to reduce superfluous operational expenses (OPEX) and lead to shorter time-to-market while also lowering capital expenditures (CAPEX).

According to Cisco Visual Networking Index [1] by 2020, over 75% of global mobile data traffic will be video content, growing from 55% in 2015, and 7 trillion video clips will be uploaded in 2020, translated into 2.5 daily video clips for every person. In other words, media services require a high consumption of computing and network resources due to stringent Quality of Service (QoS) demands, imposing challenging management scenarios of these resources.

However, emerging issues for media services and applications are beyond capacity, latency and data rate, especially in mobile environments. As representative examples the number and diverse capabilities of end-user devices, the requirements for anywhere and anytime availability and interaction between users, focusing on gaming and immersive media applications, the development and deployment complexities related to software and hardware heterogeneity, as well as the required Quality of Experience (QoE) maintenance across the network, are posing great challenges to delivering multimedia applications and services over 5G networks, as revealed by the EU 5G PPP Association [2].

Within the 3GPP RAN architecture, there is in 5G the progress towards a Next Generation Core (NGC) [3]. Our solution, is in the line of a further evolution required for true flexibility in the 5G Core, and this is achieved by the adoption of Software Defined Networking (SDN), Network Function Virtualization (NFV), Network Slicing, and Cloud RAN.

Future 5G systems are intended to step beyond traditional connectivity, orchestration and provisioning paradigms and allow for the management of distributed networking, compute and storage resources beyond proprietary equipment. Such resources will be delivered on top of convergent technologies, exposing them on demand to third party application developers and service providers, towards offering Anything as a Service (XaaS). In this way, unique capabilities will be unleashed to transmit high-quality video anywhere, store content one-hop away from the (mobile) user to be consumed anytime and use the computing capabilities of the devices or cloud services to transcode contents (even real time) and adapt them to the quality/size of the screen of any device.

This paper presents and deals with the implementation and tests of a novel Service Virtualization Platform (SVP) architecture based on NFV and SDN to facilitate the development, deployment and operation of media services on 5G networks [4]. The main advancement from the architecture perspective is the offer of an advanced cognitive management environment catering for the automated provisioning of NSs and media-related applications, and directly linking to the lifecycle management with user experience as well as optimization of infrastructure resource utilization. In addition, the preliminary testing done presented in Section V validates the advantages presented.

Another innovation offered by 5G-MEDIA SVP is the integration of serverless computing to media intensive applications in 5G networks. The serverless approach increases cost effectiveness of operation and greatly simplifies development and deployment time for application developers.

Furthermore, the 5G-MEDIA network control solution spans across the edge and core data centres of a 5G operator network, to allow the distribution of media content from central production centres to end users and vice versa. The media and network functions used in the implementation of the various services elastically cross various anchor points to personal devices, both fixed (in the home) and mobile (while the user is on the move in the 5G network). In addition, the platform provides mechanisms to flexibly adapt service operations to dynamic conditions and react upon events, for example to transparently accommodate the automated scaling of service-level resources, such as caches, transcoders and personalisation servers, enable dynamic Virtualized Network Function (VNF) placement to match migrating users and changing demand patterns, etc.

The reminder of the paper is organized as follows. Section II presents the related work, the architecture is presented in Section III, in Section IV the application scenarios are depicted, tests and results are shown in Section V and Section VI concludes the paper.

II. Related Work

There is a number of projects implementing and evaluating aspects of the ETSI NFV Management and Orchestration (MANO) [5] architecture and tools, mostly based on the MANO framework for NFV proposed by ETSI SONATA [6] and 5Gex [7] are developing MANO tools for 5G networks, extending ETSI MANO entities to accomplish their goals. CloudNFV [8] is an open platform for implementing NFV based on cloud computing and SDN and managing service-chaining structures based on network resources and characteristics. OpenMANO [9] is providing tools to interact with the compute and storage nodes in the NFV Infrastructure (NFVI) to manage VNF functions based on their performance and portability characteristics. OPNFV [10] is another open source project, aiming to validate multi-vendor, inter-operable NFV solutions. OPNFV is an open source approach of the NFVI and Virtualized Infrastructure Manager (VIM) components of the ETSI architecture, supported by many large companies worldwide. But the orientation to media optimal delivery in flexible networks, towards obtaining the best QoE in 5G networks, has not been fully achieved. One of the problems in media, is the optimal NFV placement.

Optimal placement of virtual functions in cloud computing environments is a complex decision related to the bin-packing problem which is known to be Nondeterministic Polynomial hard (NP-hard) [11]. Many heuristics have been proposed and most of them are based on greedy algorithms using simple rules [12], [13]. More complex heuristics consider grouping Virtual Machines (VMs) based on the complementarity of their workload [14], [15]. More recently, sets of VMs are scheduled on the infrastructure together, inducing additional constraints to their placement. Examples of such VM set scheduling situations are described in [16], [17]. An additional degree of freedom is introduced by VM live migration [18]. [19] proposes a set of techniques for VM rescaling, replication and live migration. Researchers have observed [20], [21] that VM performance depends on the underlying hardware and this must be accounted for an optimal placement. In literature, isolated investigations have been performed towards situations related to live migration, such as migration sequence planning and bandwidth considerations [22]. In addition, many existing studies in service deployment and selection domains have considered the trade-off between performance and deployment/transit costs [23]–[26].

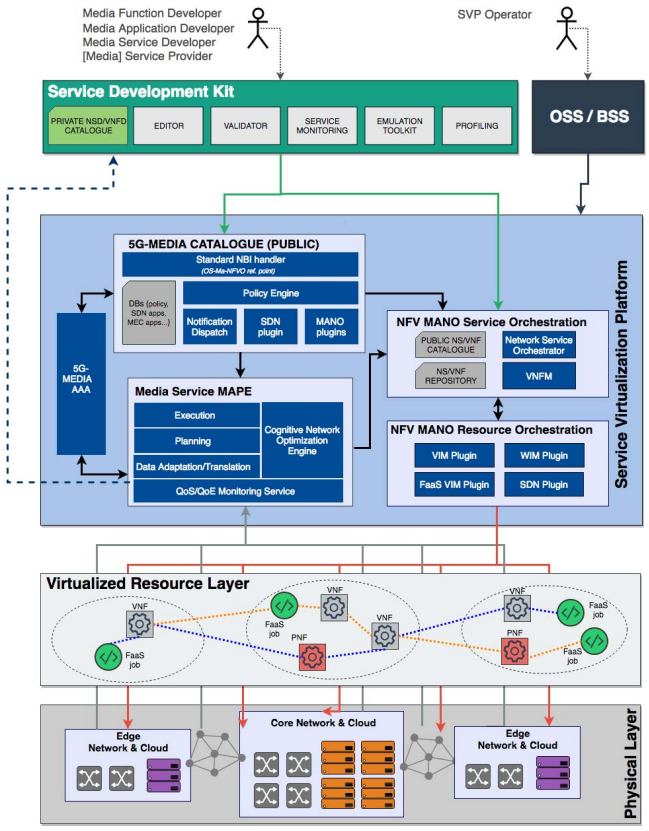

Fig. 1. High-level architecture.

III. Architecture Overview and Components

5G-MEDIA

architecture is applying SDN and NFV concepts to media applications to flexibly and dynamically embed them as virtual network functions (VFNs) within the 5G networks and cloud infrastructures. To ensure high performance levels in terms of high bandwidth and low latency, the media application functions are deployed close to traffic sources and sinks, and the 5G-MEDIA MANO function deploys smart algorithms for configuring network paths and virtual slices to deliver the required network capacity and performance levels at the network edge.

5G-MEDIA

architecture is delivering a SVP to orchestrate the deployment and scaling of the media applications, interacting automatically with the underlying network for the dynamic control of the network paths and forwarding graphs by applying Machine Learning (ML) driven optimization techniques.

VNF are software implementations of network functions that can be deployed on a NFVI. On the other hand, physical network functions (PNFs) are hardware boxes providing specific functionality, such as a broadcaster's transmission equipment. While PNFs used to be the de-facto standard for many decades, 5G networks bring VNFs as a means of flexible deployment and upgradability of network functionality compared to PNFs.

The high-level architecture description of the 5G-MEDIA platform is shown in Fig. 1.

In a top-down perspective, the architecture defines three layers of operations:

1. The development and service preparation and evaluation layer, including the Service Development SDK, which provides the means to service/apps developers and rest stakeholders to develop, emulate and deploy VNFs/NSs and access every exposed service by the SVP. The SDK provides the set of open source tools for supporting the rapid development of network applications (even on top of already existing VNFs, stored in the 5G-MEDIA VNF Repository) throughout the application/service lifecycle. It consists of proofing and packaging tools as well as emulator mechanisms to accelerate application development and provides a testing environment to be utilized prior to service deployment in the SVP cloud resources. In the 5G-MEDIA architecture, one of the major project innovations is the integration of the serverless computing approach, leveraging open source projects such as OpenWhisk [27] (while also introducing enhancements required to be the strict requirements of media applications). The main benefit from the integration of this paradigm is that developers do not need to care about the low-level details related to the infrastructure and operation specificities, thus drastically reducing development time and maintenance effort. Similarly, another innovation is the integration of unikernel packaging based on Mikelangelo EU project [28], [29], resulting in smaller footprint and safer VNFs compared to plain ISO, easily managed by the developers thanks to DevOps.
2. The SVP layer that hosts the components related to the Open Source MANO (OSM)-based MANO framework (service and resource orchestrator, Infrastructure Manager(s), Repositories, etc.), as well as components of specific purposes, i.e., the 5G-MEDIA Catalogue, the Media Service Monitor Analyse Plan and Execute (MAPE) component and the 5G-MEDIA AAA mechanism. The core component of 5G-MEDIA SVP is the MANO framework. Adopting the architectural principles of OSM, MANO functionalities are assigned over two main subcomponents in 5G-MEDIA SVP, i.e., the NFV MANO Service Orchestration (SO) and NFV MANO Resource Orchestration (RO). The SO sub-component undertakes responsibilities of NFV Orchestrator (NFVO) and VNF Manager (VNFM), while also the control of the VNF/NS Repository & Catalogue. The RO sub-component introduces a modular, customizable, and easily extensible plugin-based architectural model able to interact with multiple Wide Area Network (WAN) Infrastructure Managers (WIMs), SDNs and VIMs, including those enabling Function as a Service (FaaS) capabilities.

The 5G-MEDIA Catalogue, formally 5G Apps and Services Catalogue, is a new functional element which is designed to be NFV MANO platform-agnostic in terms of formats and syntax for NS descriptors and VNF Package information model. This catalogue uses a novel generalized and extendible format for representing NSs and VNFs, and it is capable to onboard NFV service elements from federated MANO systems (e.g.,

to complement a domain's catalogue of NSs and VNFs with items made available by other federated domains), as well as Mobile Edge Cloud (MEC) media applications and services and other virtual applications such as SDN applications, and functions implementing the FaaS paradigm (described in Section III-A and III-D).

A major innovation is the development of the Media Service MAPE component, which is composed of the Cognitive Network Optimizer (CNO), the Monitoring service, the Planning and the Execution services. The CNO Engine is taking advantage of the cognitive control principles to establish a ML-enabled optimization environment that dynamically establishes and updates the live VNF Forwarding Graphs (VNFFGs). To achieve this, it is driven by the monitoring service which aggregates various metric values of interest from every running application NS and integrated infrastructure (e.g., NFVIs). Apart from the CNO engine, these values are directly accessed through an open brokering system by the visualization tools of the SDK, as well as every internal service of SVP that may be interested in. The Planning service consists of different optimization models and caching strategies, linked with applications and tenants, supporting media and entertainment applications and their proper placement in NFVIs. Last, the Execution service triggers execution mechanisms according to the capabilities provided by OSM (i.e., scaling groups in NSDs) to enforce commands of the CNO over the integrated NFVIs and live VNFFGs.

3. The physical layer which is composed of every cloud computing, virtualization and other type of infrastructure is used to host instantiated VNFs/NSs and deliver 5G-MEDIA application services to the end users. The purpose of the Core Network and the Edge environment is three-fold: i) it provides sufficient resources to instantiate VNFs (or part of them in the microservice-based approach) that are used by multiple tenants or applications (e.g., virtual firewall), as well as application-specific helper functions/components (such as rendering and/or augmented reality servers), ii) it can be utilized to allocate resources following the network slicing concept in order to satisfy specific QoS/QoE requirements of an application or security/privacy concerns of a service provider, and iii) can be used to facilitate the deployment of legacy components and services especially those instantiated on physical/specialized hardware (that is indeed a reality in media and entertainment applications development world). Several cloud-based edge networks and cloud environments are connected to the SVP as NFVIs allowing for the instantiation of network applications closer to the user (edge computing paradigm).

In the following subsections, the role and the main (sub)components and services of 5G-MEDIA software architecture are presented in more detail.

A. MANO Framework

As already mentioned, the 5G-MEDIA architecture leverages on OSM [5] to meet the requirements of NFV/SDN network, aligned with ETSI NFV model. Thus, the two main subcomponents of MANO framework, shown in Fig. 1, are the SO and RO, respectively. Besides those, 5G-MEDIA SVP also introduces three other components, i.e., the 5G-MEDIA Catalogue, the MEDIA AAA and the Media Service MAPE. This paper focuses on the workflow and responsibilities of the Catalogue and MAPE component, which play a critical role in the SVP and are presented in the following subsections, along with the integration of FaaS framework.

FaaS is a new form of container-based Platform as a Service (PaaS) that is rapidly gaining momentum. Among the main advantages of FaaS are a higher level of abstraction offered to the application developers, significant cloud operational cost reduction thanks to a finer granularity of resource allocation, instantaneous elasticity and a finer granularity of billing.

FaaS paradigm is a compelling cost-efficient approach in use cases when workloads exhibit high peak/average resources consumption ratio and are inconsistent in the sense of comprising unscheduled/unpredictable events requiring instantaneous event handling.

The reason for that is that FaaS transfers the burden of capacity planning for these types of workloads, which is a very challenging task due to their volatility, from the application owner to the FaaS cloud platform provider. The latter can handle capacity planning task for FaaS more efficiently thanks to statistical multiplexing naturally occurring on the shared platform.

The SO sub-component performs all aspects of SO, including VNF/NS lifecycle management and end-to-end, resource-coordinated services execution in an otherwise dispersed NFV environment. In particular, the SO is responsible for constructing a service chain based on the information included in the corresponding NS Descriptor (NSD), such as the VNFs and PNFs composing the media service and the Virtual Links Descriptors (VLDs) which describe the resource requirements needed for links between VNFs, PNFs and the endpoints of the NS. In this line, the SO module supports the following operations:

1. NS/VNF instantiation: deployment of NS/VNF instances, according to the lifecycle events defined in the corresponding NSDs/VNFDs.
2. NS/VNF configuration: modifications in the configuration of NSs/VNFs through their descriptors. This can be done either prior to the instantiation of a NS/VNF or as an active process while the NS/VNF is running.
3. NS/VNF performance monitoring: Several different performance's metrics from the computing instances and the network/virtual links (e.g., bandwidth, latency etc.), as they are collected by the SVP monitoring module, are provided to the SO to track critical Key Performance Indicators (KPIs) and trigger corrective actions.
4. NS/VNF scaling: increase/decrease of the NSs/VNFs instances according to the scaling policies defined per NSDs and VNFDs. This may result to creation/termination of VNF instances or updating the virtual links over them.
5. VNFFG updates: update VNFFGs based on VNFFGDs and also recommendations provided by

the Media Service MAPE CNO engine. This may result to re-ordering VNF list and modifying traffic routes over it.

The RO is responsible for managing and coordinating resource allocations across multiple geo-distributed VIMs and multiple SDN controllers. It exposes a northbound API to communicate with the SO sub-component and provide a number of utilities for internal consumption. In line with the specifications of OSM, the Resource Orchestrator adopts a plugin programming model, which allows to add functionality without modifying or having access to its source code. Each plugin is responsible to connect the interface of the corresponding entity with the RO. The four types of plugins that are supported are:

1. *VIM plugins:* Each integrated VIM is responsible to control and manage the compute, storage, and network resources within one operator's infrastructure subdomain. In 5G-MEDIA SVP, there may be multiple VIMs where each one manages an individual infrastructure domain. Under this scope, each VIM interacts directly with the corresponding NFVI domain to deploy NSs therein and manage the available resources. Thus, every VIM implementation should be able to maintain an inventory of the physical resources and also keep track of their utilization and their map to virtualization resources. Apart from OpenStack and VMWare plugins, which are already provided by OSM release 3 and beyond, the 5G-MEDIA SVP can also interact with other VIMs.
2. *FaaS VIM plugins:* FaaS VIM plugin is a specialized VIM plugin that integrates serverless computing capabilities into the 5G-MEDIA platform. In the platform a a plugin to integrate FaaS model implemented by OpenWhisk into the SVP is available.
3. *SDN plugins:* Each integrated SDN controller undertakes the traffic/flow control throughout the underlying network elements to enable intelligent and efficient networking. 5G-MEDIA leverages on OpenDaylight, ONOS and Floodlight plugins which are available by the OSM stack [30].
4. *WIM plugins:* Each WIM plugin abstracts the interactions between multiple WANs over which the VNFs/NS may be instantiated. In particular, a WIM should implement the following functionalities related with the NFVI connectivity services: i) Path computation according to QoS input parameters; ii) Connectivity establishing over the physical network; iii) North Bound APIs to be used by the NFVO; iv) South Bound APIs/Drivers to SDN Controllers in order to configure the underlying network.

B. 5G-MEDIA Catalogue

The rational in introducing a novel 5G Apps and Services Catalogue resides in the limitations that to date we can face in operating on top of several State-of-the-Art NFVO catalogues: the high fragmentation in VNF Packages and NSDs formats and contents as well as the poor support for package versioning affects, from the DevOps perspective, the possibility of realizing a cross-platform/cross-NFVO portable offer of applications and services. At the same time, we can notice a limited support for application specific configuration and monitoring parameters, while preserving standards compliance for NSD and VNFD. Different mechanisms and procedures exist for the onboarding and management of descriptors and images for VNFs (standard and FaaS), SDN Apps, MEC Apps, etc. The catalogue brings a key set of features for suppling to above reported limitations in state of the art:

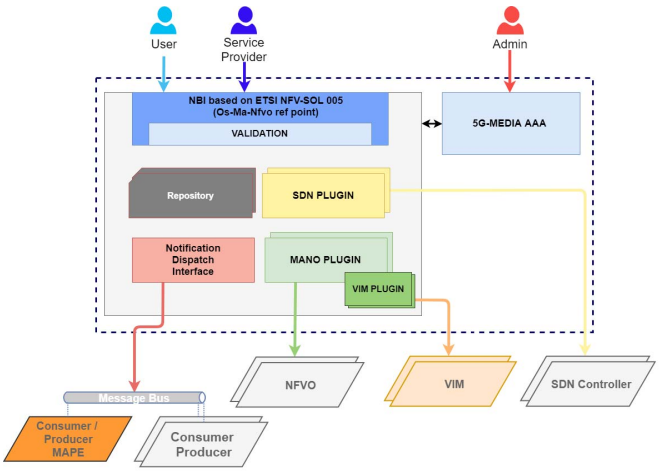

Fig. 2. 5G Apps and Services Catalogue.

- A standard, unified and extendable format for descriptors and packages: VNF packages and VNFDs based on ETSI GS NFV IFA 011 [31], ETSI GS SOL 004 [32], and ETSI GS SOL OO1 draft [33]. MEC apps based on ETSI GS MEC 010-2 [34] and NSDs based on ETSI GS NFV IFA 014 [35] and ETSI GS SOL OO1 draft. The SDN Apps modelling leverages on the outcomes from 5G-PPP SELFNET phase 1 project [36].
- A standard aligned North-Bound Interface for NSD and VNF Package Management (e.g., upload, fetch, update, delete and query) based on ETSI GS SOL 005 (Os-Ma-Nfvo reference point) [37].
- A set of MANO domain-specific translators from common to specific descriptors.
- Mechanisms for application/function images uploading with reference to different targeted VIMs.
- A Notification Engine for discovering, advertising, publishing, and validating descriptors across catalogues from different providers.

The high-level design of the 5G Apps and Services Catalogue is depicted in Fig. 2. In particular, this modular design enables a customizable deployment of the application in terms of plugins instantiation and run time plugins configuration. In fact, the southbound interface of the Catalogue is composed of different plugins capable of handling the translation of the generalized package/descriptor into the specific format expected at the underlying orchestrator (both NFV and SDN orchestrators could be supported) and actuating onboarding/management operations on the target virtualization platform. In particular, each MANO plugin includes:

- a translation module responsible for translating the generic descriptor in the format expected at the underlying MANO Service Orchestrator (e.g., packages compatible with the OSM information model specification),
- a set of VIM plugins (e.g., OpenStack plugin, OpenWhisk plugin etc.), one for each VIM in the NFVI administrated by the target MANO stack, for uploading images in the VIM images' storage,
- a MANO agent for collecting feedbacks about onboarding and instantiation operations as well as for notification about, for instance, new VIM instances or new capabilities supported by the MANO framework.

The 5G App and Service Catalogue design foresees also the implementation of a Notification Dispatch Interface for sending service and application specific notifications to a set of consumers listening on a notification bus. In 5G-MEDIA, a specific consumer on the message bus is the MAPE component, which retrieves application specific monitoring parameters used to initiate monitoring jobs once the service/application is instantiated through the MANO stack.

C. Cognitive Networking: Media Service MAPE

The MEDIA service MAPE (Monitoring, Analysis, Planning and Execution [38] component has been designed to deliver QoS-based control and management functionality for 5G media services. This goes beyond the orchestration logic expected of standard MANO components and aims to provide a set of tools and algorithms for the automated optimization of network and computational resources, provisioning of dynamic VNFFG adaptation, NS monitoring and QoS/QoE guarantees. The main objectives of the Media Service MAPE are:

- To collect and store metrics about the status of infrastructure resources and the performance and behaviour of NSs and media applications.
- To organize and harmonize collected metrics under a common data model.
- To integrate ML and resource planning algorithms to predict state and thus optimize the media applications and the NFVI resources.
- To implement deployment and scaling directives to MANO components to optimize resource management, network performance and enforce QoE guarantees.

The reference control model for resource optimization and dynamic VNFFG adaptation in 5G-MEDIA is shown in Fig. 3 and is based on the following five steps.

1. Monitoring and data collection: Information from running VNFs, the NFVI and networking environment is gathered by the data collector.
2. Analysis and prediction: Monitored data is processed by the ML engine to predict/forecast future trends in user demand, network conditions and resource availability. The intelligent forecasts become the input for the policy/optimization component.
3. Resource allocation/planning: By deploying a range of optimization techniques from traditional optimization approaches based on linear programming or heuristics to deep learning Artificial Intelligence (AI) algorithms, this component is responsible for devising a plan, based on policies, to efficiently use resources in order to achieve system goals. The 5G-MEDIA SVP implementation foresees four different cases for optimization algorithms with different targets:
   a. Service placement optimization to determine which NFVI instance/edge node should house each VNF for a NS by trading-off cost with performance of the network and computational infrastructure. This can run at various timescales, including initial NS deployment and ongoing reconfiguration to migrate existing VNFs, instantiate new VNFs, undertake service scaling as demand patterns change.
   b. VNFFG optimization to determine which instances of VNFs should be interconnected to meet performance and cost objectives for specific user session requests. This can be undertaken at initial session establishment as well as for the optimization of already running sessions/VNFFG instances.
   c. Infrastructure adaptation to overcome streaming difficulties, e.g., to reserve network capacity, allocate greater computational capacity for stream processing, establish expedited paths or reroute flows to avoid congested parts of the network.
   d. Application-specific adaptation and intelligent network-wide congestion avoidance, for example to configure the capturing or transcoding of 3D models to defined quality levels to match dynamically varying network throughput capabilities and available processing capacity along the NFVI nodes and clusters implementing the VNFFG instance.
4. Execution: The MANO SO components send instructions to the RO components to instantiate and configure the VNFs.
5. The VNFs in the underlying NFVIs meet the new optimization objectives determined by the steps 3 and 4.

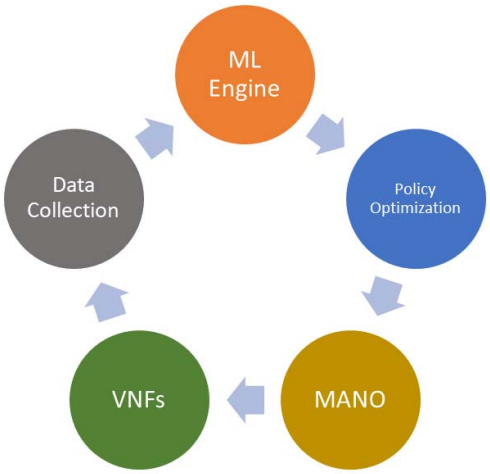

Fig. 3. Cognitive Network Optimization workflow in the Media Services MAPE component.

One of the main components of the CNO is the Policy/Optimization component, shown in Fig. 4. A range of

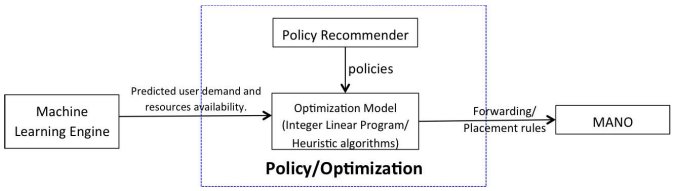

Fig. 4. Policy/optimization component.

options exist for implementing the optimization algorithms. For instance, Integer Linear Programs can be used for small input datasets in which an optimal solution can be obtained within an acceptable time. Alternatively, suitable heuristic algorithms can be used to find close-to-optimal solutions for larger input datasets. A third option is to implement the optimization decisions using machine/deep/reinforcement learning techniques rather than using ML solely as a means to forecast demand, which is then input into traditional optimization functions. The optimization model is programmable to follow policies defined by the service provider. For example, the service provider can determine the weight that the optimization algorithm gives to reducing costs versus improving performance, or it may define the maximum cost budget for any particular solution, or the maximum latency acceptable for its users.

It should be noted that the results presented in Section V-C on ML for anomaly detection are related to the traffic prediction part of the CNO. Results on the optimization of service placement and resource allocation in the context of the wider 5G-MEDIA platform will be reported in future publications.

D. FaaS Framework

In this paper we propose using FaaS to orchestrate media VNFs. This approach is well aligned with the on-going cloud-native transformation in NFV, which is powered by advances in container and microservices technologies [39], [40]. In [29] FaaS has been proposed to be part of MEC in the context of massive Internet of Things (IoT) applications. To the best of our knowledge, FaaS was not previously proposed as a mechanism for VNF orchestration either for general network VNFs or media specific ones.

FaaS model elevates the level of abstraction for the VNF developer. While, at run time, serverless functions execute as containers, developers can abstract away the details of containers preparation and management and focus only on the VNF code.

Fig. 5 depicts a software architecture for reference implementation of integration of FaaS in 5G-MEDIA SVP.

The architecture is comprised of three main building blocks: (a) FaaS Plugin, (b) FaaS Framework, (c) Container Orchestration Engine, and (d) Infrastructure as a Service (IaaS) layer.

The requirements underpinning the FaaS framework integration with the 5G-MEDIA platform can be summarized as follows.

- FaaS VNF lifecycle management (on-boarding, instantiation, monitoring, and deletion) should be fully aligned with that of the non-FaaS VNFs and facilitated through the same MANO stack;
- Since FaaS frameworks rapidly proliferate, the architecture should provide for extensibility, i.e., it should be easy to incorporate new FaaS frameworks;
- FaaS VNFs should be: discoverable and accessible over the public network similarly to non-FaaS VNFs; able to access services over the public network (including other VNFs); and FaaS and non-FaaS VNFs can be chained (both in a single DC and across DCs);
- FaaS VNFs should be able to exploit special hardware, such as GPUs;
- It should be possible to influence FaaS placement through collocation and anti-collocation constraints;
- FaaS framework deployment and operation should be independent of the underlying cloud virtualization technology
- Minimal changes to the ETSI MANO workflows.

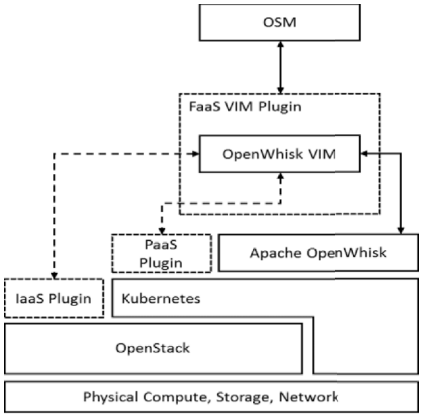

Fig. 5. Integration between OSM and OpenWhisk.

The specific choice of open source technologies for implementing this approach is made based on their maturity and popularity in the industry.

5G-MEDIA

SVP implements a new plugin for Apache OpenWhisk, which allows to instantiate functions that implement VNFs, in response to requests by OSM. The VNFs are defined as OpenWhisk *actions* prior to management flows of OSM can be enacted. This is similar to pre-population of other types of VIMs, e.g., Open Stack, with the VNF VM images. In case of OpenWhisk, the image is a combination of the code and the metadata describing it, which is stored in the OpenWhisk database.

When an OpenWhisk action is being invoked, the action code is being automatically injected into a container that corresponds to the language environment. The container is executed on top of a container orchestrator. In our reference implementation, Kubernetes (also referred to as K8s) is used as container orchestrator. K8s executes containers within pods. Using K8s abstraction of a Service, OpenWhisk actions executing in 5G-MEDIA can be discovered by other actions and they can communicate via the network (a requirement which is unique for the media functions) using out of the box K8s Flannel networking.

Also, K8s allows to influence placement of pods on server in K8s node via scheduling policies that can be specified for the pods. We use this functionality of K8s to allow proper usage of GPUs by media intensive VNFs (this is another requirement pertinent to many media intensive VNFs).

K8s can run either on bare metal server or on VMs. In 5G-MEDIA reference implementation, Open Stack is used as the IaaS layer.

IV. Application Scenarios: Modules and Functionalities

A. Use Case 1: Tele-Immersive Media Application Scenario

With 5G set to enable new forms of immersive media, the first scenario implemented under the proposed platform's architecture is a tele-immersive (TI) media application [41], [42]. In this context, two users are remotely interacting in a gaming context with each other via their real-time 3D reconstructed replicas which are transferred as 3D multimedia streams over the network, allowing for unrestricted free viewpoint rendering. In addition, their interactive session can be spectated by an arbitrary group of users. However, since novel applications are usually volatile in terms of traffic, a traditional service design and deployment model would severely lack OPEX efficiency. This scenario is thus, an ideal candidate for applying the serverless approach. We design a micro-service oriented real-time transcoding service where a virtual 3D media transcoder VNF (vT3D) is placed on the most proximal edge to each user. These vT3D VNFs are responsible to encode incoming traffic in multiple qualities which can then be requested by both types of remote users (interacting users and spectators) according to their adaptation logics.

Each TI session is deployed on demand using OSM, and therefore we require a light-weight virtualization technology – i.e., containers – to accommodate for the fast deployment times required to keep session instantiation response time low, and thus, increase QoE. At the same time, in this way we can utilize the wider availability of resources to deploy VNFs at the appropriate edges to minimize core traffic and improve latency.

The FaaS plugin spawns the micro-service components (vT3D) upon each session's instantiation. Therefore, the volatility of the sessions' uptime is perfectly matched by the flexibility and elasticity offered by the SVP through its plugged in serverless framework. Given that our TI media application is interactive, it can further utilize FaaS by exploiting specific in-game interactions to trigger appropriate media functions. In this way, our scenario foresees the automatic generation of replay clips at certain game events. These would be initiated through OSM and deployed via the FaaS plugin to the core cloud. A diagram depicting this scenario is presented in Fig. 6.

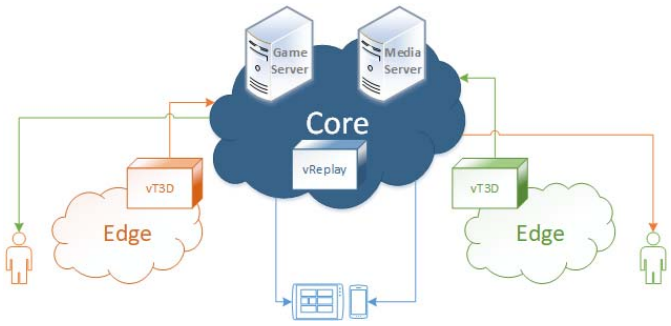

Fig. 6. The tele-immersive game media service architecture.

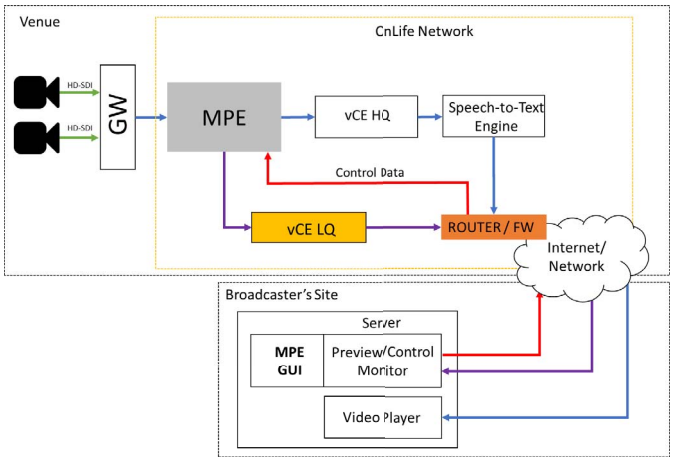

Fig. 7. Remote Production NS topology.

B. Use Case 2: Remote and Smart Production in Broadcasting Scenario

This use case aims to demonstrate the benefits that the advancements in 5G technology, bring to professional remote broadcast productions [43]. In this domain, when the production of a live event needs to be done, a gateway (GW) is set up in order to convert the video signal from Serial Digital Interface (SDI) to IP. After that, the broadcaster uses the 5G-MEDIA platform to connect a virtual Media Process Engine (MPE) serving as video switcher, virtual Compression Engines (vCE) serving as encoders and a Speech-to-Text Engine (S2T). All these VNFs are onboarded through the 5G-MEDIA Apps and Service Catalogue where the VNFFG is defined in order to compose the NS that allows the broadcaster to perform the remote production, saving personal and technical costs for the broadcaster. This NS is presented in Fig. 7.

Once this NS is defined, it is instantiated by using the NFVO and VIMs of the 5G-MEDIA SVP. Finally, since the end user in this use case is the broadcaster, an optimal QoE needs to be assured. For that reason, the platform uses the 5G-MEDIA MAPE, in order to monitor and optimize the NS, aiming to that QoE optimization.

C. Use Case 3: Ultra-High Definition (UHD) Over Content Distribution Networks (CDN)

This use cases targets the UHD media delivery over virtualized content distribution networks [44]. Leveraging on 5G technologies, we aim to develop a virtual CDN (vCDN) solution capable of meeting the needs of the increasing media industry, where there's a high demand for services

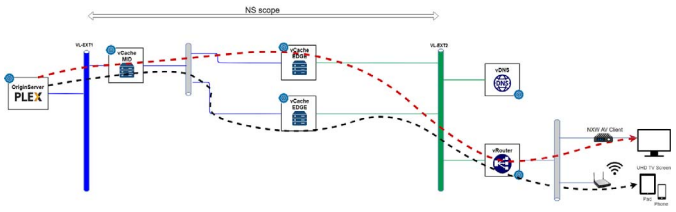

Fig. 8. vCDN NS topology.

capable of distributing different types of media contents, with a high volume of data depending on the media quality, to an heterogenous set of end-devices (e.g., phones, pads and TV screens) connected to the network. The simplified version for the vCDN NS topology is depicted in Fig. 8.

The vCDN NS is composed by a vCache hierarchy bind to two external networks, one (on the left) interconnecting the first level of vCaches with the UHD-capable Origin Streaming Server and the other one (on the right) that allows the end-users to access the service through a load balancing mechanism. The first implementation of Use Case 3 (UC3) targets the scenario "My screen follows me" [45], where we envision the daily life of a typical family (example): Diego, the father, is a business man, Mariana, the mother is a musician and their child Daniel is a student. For instance, Mariana would like to stream the opera she is working on while traveling from home to the Opera House, then switching the streaming of the media content of her personal library from the TV to her phone and continue listening to it while moving in urban environment. For offering a proper QoE to the end-user, with a zero-perceived interruption of the streaming service, the instantiated vCDN must adapt its vCaches hierarchy. The adaptation of the streaming hierarchy is performed by the 5G-MEDIA MAPE and could be triggered by several events that even potentially could affect the streaming quality [46]. An initial evaluation of ML-based traffic forecasting and anomaly detection algorithms used to trigger scale out actions on the vCache hierarchy is in Section V-C. The components from the SVP involved in the vCDN SO are:

- **5G Apps & Services Catalogue**, for NSD and VNF Packages onboarding along with Monitoring Descriptors.
- **NFVO (OpenSource MANO) + VIMs**, for the service instantiation and service lifecycle management operations.
- **5G-MEDIA MAPE**, for NFVI and application level monitoring. The CNO module is responsible for data analysis and processing in order to adapt and optimize the vCDN service.

V. Testing and Results

A. FaaS Performance Evaluation

In this subsection, we describe our preliminary evaluation of the FaaS integration architecture. The application used in our evaluation is depicted in Fig. 6. The testbed comprises a five node K8s cluster. Three nodes are VMs deployed on Open Stack Queens release. The VMs run Ubuntu 16.06 with 2 vCPUs, 4GB of main memory and 40GB of hard disk. One of the VMs hosts K8s Master, another VM hosts an all-in-one OpenWhisk installation, and the third VM executes additional supporting services, such as a service that offloads an OpenWhisk action to K8s. The other two nodes are Desktops with NVIDIA GPU cards, one Intel Core i7, 32GB RAM, and 2TB HDD. The nodes run Ubuntu 16.06.

TABLE I
Invocation Latency of the OpenWhisk Pluging vs Direct K8s Invocation

Statistic	K8s Direct invocation	OpenWhisk Plugin
Mean	7.07 s	7.8 s
Confidence Interval for mean (at 0.05)	±0.2045 s	±0.2778 s
Minimum	6.492 s	6.351 s
Maximum	8.161 s	8.868 s
Standard Deviation	0.495 s	0.673 s
Kurtosis	0.044	-0.558
Skewness	1.07	-0.445

The goal of our preliminary evaluation study was to quantify the performance cost of a higher abstraction offered by FaaS vs directly using containers on top of container orchestrator.

To that end, we are interested in measuring the latency of VNF instantiation (we used the vT3DTranscoder VNF of UC1 for our experiments) via our OpenWhisk plugin as opposed to direct instantiation on K8s.

We perform 30 experiments of each of the two types. In each experiment, we instantiate a VNF and measure the time that elapses between the invocation request and pod ready state reported by K8s Master in response to a state polling request that we run every 0.2 seconds against the K8s Master. The results are summarized in Table I.

As expected, OpenWhisk introduces some overhead on top of K8s. While performance degradation of instantiation amounts to 10% on average, the absolute difference is less than one second, which makes it virtually non-observable to the human participating in the gaming session. It should be noted that the gaming session itself goes on for a few minutes. Hence, a sub-second delay during instantiation is negligible. During the session, there is no overhead added by OpenWhisk plugin as compared to raw K8s. The measurements behave well, with low sample variance, moderate skewness, and the low value of kurtosis indicates low propensity for tail outliers. This preliminary result is encouraging, because it means that advantages of FaaS programming model for VNF orchestration can be leveraged with very low performance overhead for media intensive applications.

B. Media-Based QoE for Service Re-Configuration. Preliminary Results

5G advances in terms of content distribution have led to an increase of the need of integrating media quality assessment as a primary feedback for video services evaluation. At the same time, quality considerations have become a lot more challenging as more complexity both for content and networks.

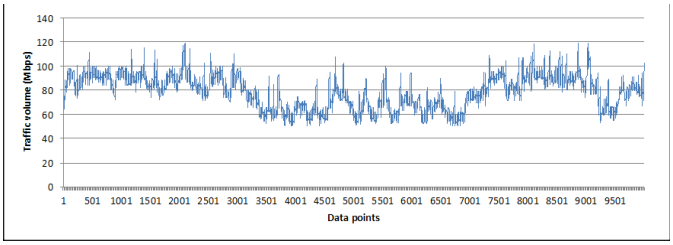

Fig. 9. Example aggregate traffic trace used for ML training.

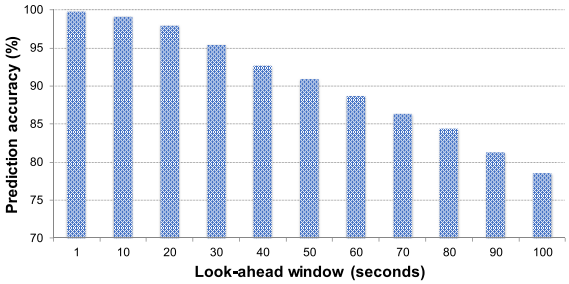

Fig. 10. Prediction accuracy versus look-ahead window.

Opposite to classic QoS systems that are based on network assessment, media-based QoE takes into account not only system performance parameters but content objective quality.

5G-Media QoE probe has proved itself valuable to provide with no-reference quality measures of video services as offered in the proposed use cases. Image artefacts, content complexity, parametric coding information, and other intrinsic data extracted directly from the content are gathered to provide a Mean Opinion Score (MOS)-style score regarding the quality of the video service [47]. This information was previously completely blind for the network management system. Thus, bad quality content services could be offered without any evidence for the system, unable to detect and correct this problem of great impact on the user. The reaction time can be set below 15 s, with a complete reconfiguration of the SVP, maximising the QoE, in 20-40s.

By exploring the accuracy requirements of potential uses as well as evaluation criteria, the probe sets the stage to make substantial future improvements to the challenging problem of No-Reference (NR) quality estimation for upcoming 5G content service deployments [48], [49].

C. ML Approach to Anomaly Detection in the MAPE Loop

The CDN NS depicted in Fig. 8 was used as the basis for the evaluation of the CNO's ML algorithms for anomaly detection and traffic forecasting. The scenario is one where anomalous traffic, a flash-crowd event, for example, causes congestion on the network between vCaches and the users, which may cause performance degradation for the delivered video, reducing QoE for the users. The solution to which is the triggering of a scale-out operation by the MAPE component to deploy additional vCaches available over non-congested network segments.

Two sources of real traffic datasets were selected for the initial training and testing of the ML algorithms: from transit links maintained by the MAWI Working Group of the WIDE Project [50], and from backbone links by CAIDA [51].

For controlled testing and tuning of the algorithms we used synthetically generated traces for the anomalous traffic, emulating flash crowd events. The real background traces were aggregated with the synthetic anomalies to create the training and testing data of the algorithm. An example trace of 10000 data points for offered load is show in Fig. 9, where it can be seen that there are some periods where the total offered load is greater than the link capacity of 100Mbit/s. In our model we defined congestion as five or more continuous data points greater than 100 Mbit/s (this value is a parameter of the ML algorithm and can be set at run time).

We designed a ML algorithm with the objective of predicting when congestion will happen in the near future in order to trigger the scale-out of the vCaches. We used a fully connected neural network with one hidden layer of ten nodes, implemented with TensorFlow [52] and Keras [53]. We use Adam as an optimization algorithm to update network weights iteratively based on training data and ReLU (Rectified Linear Unit) as the activation function.

The model was trained using the aggregate traffic trace shown in Fig. 9. We tested the model using a traffic trace generated from a separate 10000 data point sample from the real traffic traces aggregated with synthetic anomalous load spikes. The inputs to the ML model was a moving window of ten data points: the current measured load and the prior nine measurements.

Each training input consisted of the moving window of load values plus a label of whether there was a congestion event within a defined look-ahead window immediately following the current time. The purpose was to train the ML algorithm to identify early characteristics of the traffic anomalies and to identify whether the link would be congested within the look-ahead period.

We found that this type of training data contained redundant information concerning the average load values, which prevented the ML model from identifying the key characteristics of the anomaly traffic and giving accurate predictions within a reasonable period of training. The test set accuracy we measured was between 75% and 79%. For this reason, we decided to use training data with delta values with the first input being the initial absolute traffic volume value and the following nine inputs being delta values of offered load. The insight being that the ML algorithm may be better at identifying the characteristics of anomaly traffic patterns when explicit delta values between data points are presented as input. The results with this training data over the same training period were significantly improved, giving a test set accuracy of 93%.

The graph in Fig. 10 shows how the prediction accuracy of the ML algorithm varies with the length of the look-ahead period. In terms of system performance, when running on a standard laptop the training period took around 8 minutes for 201 epochs. Prediction time once the model has been trained is virtually instantaneous.

The results so far have been limited in scope to detecting anomalies on a single link. Extensions currently under study

will make use of metrics from multiple network links as well as the computational load on the vCaches. Inputs from a wider variety of data sources will allow ML algorithms to detect conditions that arise through a combination of load and utilisation metrics collected from a distributed set of resources. Although this will enable more general forecasts to be made, this will be at the expense of longer and more complex training phases. The use of reinforcement learning and unsupervised learning will be further explored for such cases.

VI. Conclusion

This paper presents how an ETSI NFV compliant architecture enables new possibilities to facilitate the development, deployment and operation of media services on top of the upcoming 5G networks, leveraging cutting-edge technologies and computing paradigm like: i) FaaS as a new form of container based PaaS able to provide significant cloud operational cost reduction thanks to a finer granularity of resource allocation, instantaneous elasticity and a finer granularity of billing; ii) ML-driven optimization techniques for the optimal allocation and operation of media NSs and iii) a catalogue as a functional element designed to be NFV MANO platform-agnostic in terms of formats and syntax for NS descriptors and VNF Package information model. The proposed architecture is being validated against three media use cases: an immersive Virtual Reality 3D gaming application, the remote production of broadcast content incorporating user generated contents, and dynamically adaptive CDNs for the intelligent distribution of UHD content and first results have been provided in this paper. Extensions to the three use cases will be implemented and validated till 2019-Q4, spanning the applications to include support to mobility, higher data rates and instantiation of the needed NSs across different domains

Acknowledgment

The authors would like to thank to all members of the consortium of the 5G-MEDIA project, which has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 761699.

References

[1] *Cisco Visual Networking Index: Forecast and Methodology 2016–2021*, Cisco, San Jose, CA, USA, 2017.

[2] (Jan. 19, 2016). *5G and Media & Entertainment Whitepaper*. Accessed Oct. 2018. [Online]. Available: https://5g-ppp.eu/wp-content/uploads/2016/02/5G-PPP-White-Paper-on-Media-Entertainment-Vertical-Sector.pdf

[3] M. Shafi *et al.*, "5G: A tutorial overview of standards, trials, challenges, deployment, and practice," *IEEE J. Sel. Areas Commun.*, vol. 35, no. 6, pp. 1201–1221, Jun. 2017.

[4] S. Rizou *et al.*, "A service platform architecture enabling programmable edge-to-cloud virtualization for the 5G media industry," in *Proc. IEEE Int. Symp. Broadband Multimedia Syst. Broadcast. (BMSB)* Valencia, Spain, Jun. 2018, pp. 1–6.

[5] *Network Functions Virtualisation (NFV); Management and Orchestration, v1.1.1*, ETSI Standard ISG GS NFV-MAN 001, Dec. 2018.

[6] H2020, 5G-PPP, SONATA. *Service Development and Orchestration in 5G Virtualized Networks*. Accessed: Jan. 15, 2019. [Online]. Available: http://www.sonata-nfv.eu

[7] *5Gex Project, H2020*. Accessed: Jan. 15, 2019. [Online]. Available: http://www.5gex.eu/

[8] *CloudNFV*. Accessed: Jan. 15, 2019. [Online]. Available: http://www.cloudnfv.com/WhitePaper.pdf

[9] D. R. Lopez, "OpenMANO: The dataplane ready open source NFV MANO stack," in *Proc. IETF 92 Meeting*, Dallas, TX, USA, Mar. 2015, pp. 1–28.

[10] *OPNFV*. Accessed: Jan. 15, 2019. [Online]. Available: https://www.opnfv.org/

[11] E. G. Coffman, Jr., M. R. Garey, and D. Johnson, *Approximation Algorithms for NP-Hard Problems*. Boston, MA, USA: PWS, 1997, pp. 46–93.

[12] R. Panigraphy, K. Talwar, L. Uyeda, and U. Wieder, "Heuristics for vector bin packing," Microsoft Res., Bengaluru, India, Rep. 1, 2011.

[13] D. Wilcox, A. McNabb, and K. Seppi, "Solving virtual machine packing with a reordering grouping genetic algorithm," in *Proc. IEEE Congr. Evol. Comput. (CEC)*, New Orleans, LA, USA, 2011, pp. 362–369.

[14] B. Viswanathan, A. Verma, and S. Dutta, "CloudMap: Workload-aware placement in private heterogeneous clouds," in *Proc. 13th IEEE/IFIP Netw. Oper. Manag. Symp. (NOMS)*, 2012, pp. 9–16.

[15] X. Meng *et al.*, "Efficient resource provisioning in compute clouds via VM multiplexing," in *Proc. 7th Int. Conf. Autom. Comput. (ICAC)*, 2010, pp. 11–20.

[16] L. Shi, B. Butler, D. Botvich, and B. Jennings, "Provisioning of requests for virtual machine sets with placement constraints in IaaS clouds," in *Proc. 12th IFIP/IEEE Int. Symp. Integr. Netw. Manag. (IM)*, Ghent, Belgium, 2013, pp. 499–505.

[17] K. Konstanteli, T. Cucinotta, K. Psychas, and T. Varvarigou, "Admission control for elastic cloud services," in *Proc. 5th IEEE Int. Conf. Cloud Comput. (CLOUD)*, Honolulu, HI, USA, 2012, pp. 41–48.

[18] C. Clark *et al.*, "Live migration of virtual machines," in *Proc. 2nd Conf. Symp. Netw. Syst. Design Implement. (NSDI)*, 2005, pp. 273–286.

[19] U. Sharma, P. Shenoy, S. Sahu, and A. Shaikh, "Kingfisher: Cost-aware elasticity in the cloud," in *Proc. IEEE Int. Conf. Comput. Commun. (Infocom)*, Shanghai, China, 2011, pp. 206–210.

[20] J. Dejun, G. Pierre, and C.-H. Chi, "Resource provisioning of Web applications in heterogeneous clouds," in *Proc. 2nd USENIX Conf. Web Appl. Develop. (WebApps)*, Portland, OR, USA, 2011, pp. 5–15.

[21] J. Schad, J. Dittrich, and J.-A. Quian-Ruiz, "Runtime measurements in the cloud: Observing, analyzing, and reducing variance," *Proc. VLDB Endowment*, vol. 3, no. 1–2, pp. 460–471, 2010.

[22] S. Ghorbani and M. Caesar, "Walk the line: Consistent network updates with bandwidth guarantees," in *Proc. 1st Workshop Hot Topics Softw. Defined Netw. (HotSDN)*, New York, NY, USA, 2012, pp. 67–72.

[23] B. Jennings and R. Stadler, "Resource management in clouds: Survey and research challenges," *J. Netw. Syst. Manag.*, vol. 23, no. 3, pp. 567–619, 2015.

[24] P. Wendell, J. Jiang, M. Freedman, and J. Rexford, "DONAR: Decentralized server selection for cloud services," in *Proc. ACM SIGCOMM Conf.*, New Delhi, India, 2010, pp. 231–242.

[25] H. Xu and B. Li, "Joint request mapping and response routing for geo-distributed cloud services," in *Proc. IEEE INFOCOM*, Turin, Italy, 2013, pp. 854–862.

[26] Z. Zhang *et al.*, "Optimizing cost and performance in online service provider networks," in *Proc. 7th USENIX Conf. Netw. Syst. Design Implement. (NSDI)*, San Jose, CA, USA, 2010, p. 3.

[27] *Apache OpenWhisk*. Accessed: Jan. 15, 2019. [Online]. Available: https://openwhisk.apache.org/documentation.html

[28] *MIKELANGELO Project*. Accessed: Jan. 15, 2019. [Online]. Available: https://www.mikelangelo-project.eu/technology/unikernel-application-management/

[29] S. Kekki *et al.*, "MEC in 5G networks," Sophia Antipolis, France, ETSI, White Paper, Jun. 2018.

[30] ETSI ISG. *Open Source MANO (OSM)*. Accessed: Jan. 15, 2019. [Online]. Available: http://www.osm.etsi.org/

[31] *Management and Orchestration; VNF Descriptor and Packaging Specification, v3.1.1*, ETSI Standard ISG GS NFV-IFA 011, Aug. 2018.

[32] *Protocols and Data Models; VNF Package Specification, v2.4.1*, ETSI Standard ISG GS NFV-SOL 004, Feb. 2018.

[33] *Protocols and Data Models; NFV Descriptors Based on TOSCA Specification, v0.0.9*, ETSI Standard ISG GS NFV-SOL 001, Jun. 2018.

[34] *Mobile Edge Management; Part 2: Application Lifecycle, Rules and Requirements Management, v1.1.1*, ETSI Standard ISG GS MEC 010-2, Jul. 2017.

[35] *Management and Orchestration; Network Service Templates Specification, v3.1.1*, ETSI Standard ISG GS NFV-IFA 014, Aug. 2018.

[36] "Deliverable D3.1: Report and prototype implementation of the NFV & SDN repository," SELFNET Consortium, Brussels, Belgium, Rep. 3.1, Sep. 2016.

[37] *Protocols and Data Models; Restful Protocols Specification for the Os-Ma-NFVO Reference Point, v2.4.1*, ETSI Standard ISG GS NFV-SOL 005, Feb. 2018.
[38] M. Maurer, I. Breskovic, V. C. Emeakaroha, and I. Brandic, "Revealing the MAPE loop for the autonomic management of cloud infrastructures," in *Proc. IEEE Symp. Comput. Commun. (ISCC)*, Kerkyra, Greece, 2011, pp. 147–152.
[39] "From webscale to telco the cloud native journey," 5G-PPP Softw. Netw., Brussels, Belgium, Working Group, Jul. 2018.
[40] Y. L. Chen and A. Bernstein, "Bridging the gap between ETSI-NFV and cloud native architecture," in *Proc. SCTE/ISBE Fall Tech. Forum*, Denver, CO, USA, Oct. 2017, pp. 1–27.
[41] A. Doumanoglou *et al.*, "A system architecture for live immersive 3D-media transcoding over 5G networks," in *Proc. IEEE Int. Symp. Broadband Multimedia Syst. Broadcast. (BMSB)*, Valencia, Spain, Jun. 2018, pp. 11–15.
[42] N. Zioulis *et al.*, "3D tele-immersion platform for interactive immersive experiences between remote users," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Phoenix, AZ, USA, Sep. 2016, pp. 365–369.
[43] M. Keltsch *et al.*, "Remote production and mobile contribution over 5G networks: Scenarios, requirements and approaches for broadcast quality media streaming," in *Proc. IEEE Int. Symp. Broadband Multimedia Syst. Broadcast. (BMSB)*, Valencia, Spain, Jun. 2018, pp. 1–7.
[44] G. Carrozzo *et al.*, "Virtual CDNs over 5G networks: Scenarios and requirements for ultra-high definition media distribution," in *Proc. IEEE Int. Symp. Broadband Multimedia Syst. Broadcast. (BMSB)*, Valencia, Spain, Jun. 2018, pp. 1–5.
[45] 5G-MEDIA Consortium. *Deliverable D2.2: 5G-MEDIA Requirements and Use Case Refinement*. Accessed: Jan. 15, 2019. [Online]. Available: http://www.5gmedia.eu/cms/wp-content/uploads/2018/07/5G-MEDIA-D2.2-v1.0.pdf
[46] 5G-MEDIA Consortium. *Deliverable D6.1: 5G-MEDIA Use Case Scenarios and Testbed*. Accessed: Jan. 15, 2019. [Online]. Available: http://www.5gmedia.eu/cms/wp-content/uploads/2018/09/5G-MEDIA_D6.1_5G-MEDIA-Use-Case-Scenarios-and-Testbed_v1.0_final.pdf
[47] J. P. López *et al.*, "Virtualized module for distributed quality assessment applied to video streaming in 5G networks environments," in *Proc. IEEE Int. Symp. Broadband Multimedia Syst. Broadcast. (BMSB)*, Valencia, Spain, Jun. 2018, pp. 1–6.
[48] J. J. Giménez, P. Renka, S. Elliott, D. Vargas, and D. Gómez-Barquero, "Enhanced TV delivery with eMBMS: Coverage evaluation for rooftop reception," in *Proc. IEEE Int. Symp. Broadband Multimedia Syst. Broadcast. (BMSB)*, Valencia, Spain, Jun. 2018, pp. 1–5.
[49] W. Guo, M. Fuentes, L. Christodoulou, and B. Mouhouche, "Roads to multimedia broadcast multicast services in 5G new radio," in *Proc. IEEE Int. Symp. Broadband Multimedia Syst. Broadcast. (BMSB)*, Valencia, Spain, Jun. 2018, pp. 1–5.
[50] MAWI WG. *WIDE Project*. Accessed: Jan. 15, 2019. [Online]. Available: http://mawi.wide.ad.jp/mawi/
[51] *CAIDA*. Accessed: Jan. 15, 2019. [Online]. Available: http://www.caida.org/data/
[52] *TensorFlow*. Accessed: Jan. 15, 2019. [Online]. Available: https://www.tensorflow.org/
[53] *Keras Library*. Accessed: Jan. 15, 2019. [Online]. Available: https://keras.io

Authors' photographs and biographies not available at the time of publication.