Monitoring and Analytics for the Optimisation of Cloud Enabled Small Cells

J. Pérez-Romero[(1)], V. Riccobene[(2)], F. Schmidt[(3)], O. Sallent[(1)], E. Jimeno[(4)], J. Fernández[(5)], A. Flizikowski[(6)], I. Giannoulakis[(7)], E. Kafetzakis[(8)]

[(1)]*Universitat Politècnica de Catalunya (UPC),* Spain, [(2)]*Intel Research and Development,* Ireland, [(3)]*NEC Laboratories Europe,* Germany [(4)]*ATOS,* Spain, [(5)]*i2CAT,* Spain, [(6)] *IS Wireless,* Poland, [(7)] *National Center for Scientific Research Demokritos,* Greece, [(8)]*ORION*, Greece

***Abstract*—With the wide range of application scenarios and associated requirements expected for the Fifth Generation (5G) networks, the increased complexity in the management of the network will require smart mechanisms able to collect a multitude of different metrics and to analyse them to extract the knowledge that can drive an efficient decision making. For this reason, future networks are expected to incorporate telemetry and analytics tools able to support the different management procedures. In this context, this paper presents the monitoring and analytics framework that is being developed by the 5G ESSENCE project for optimising the provision of Small Cell as a Service in multi-tenant scenarios. The paper discusses the different components of the framework and illustrates its capabilities for the optimisation of the cloud resource allocation and the Radio Access Network.**

Keywords—5G; Monitoring; Telemetry; Analytics; Small Cells as a Service; NFV

I. INTRODUCTION

Fifth Generation (5G) systems are intended to simultaneously support a wide range of application scenarios associated with multiple market segments (e.g. automotive, utilities, smart cities, high-tech manufacturing) as well as new business models (e.g. neutral host providers, network as a service, etc.) [1]. Therefore, 5G networks will be highly heterogeneous at different levels (e.g. multiple Radio Access Technologies [RAT], multiple spectrum bands, multiple types of devices and services, etc.) and will make extensive use of Network Function Virtualization (NFV) technologies [2]. NFV allows the software implementation of network functions running on general purpose computing/storage resources and thus offers an inherent flexibility to add new functionalities, extend, upgrade or evolve existing ones thus adapting to the 5G heterogeneous scenarios.

With all the above ingredients, the management and optimisation of 5G networks will increase in complexity with respect to prior systems, thus requiring a higher degree of automation in these procedures. In this respect, the introduction of monitoring and analytics becomes a fundamental tool for efficiently managing the network [3]. Monitoring provides the ability to collect information about resources and related counters to monitor the runtime status of the compute, network and storage infrastructures of the network and the performance of the services. Analytics, in turn, generate actionable insights from the data collected by the monitoring system and allow gaining in-depth and detailed knowledge from this data, understanding hidden patterns, data structures and relationships and therefore supporting the different decision making processes. While the use of telemetry for infrastructure and network monitoring is a well-established discipline, its applicability for the new 5G use cases requires an evolution of both telemetry architectures and analytics approaches to cope with the explosion of monitoring data in terms of volume, velocity and variety [4].

Among the current activities towards the development of 5G systems, and within the scope of the second phase of the 5G infrastructure Public Private Partnership (5G-PPP), the 5G ESSENCE project [5] addresses the provision of Small Cells as a Service (SCaaS) to facilitate a third-party provisioning of radio access capacity to mobile network operators in dense localised areas. The project proposes a highly flexible and scalable platform relying on Cloud Enabled Small Cells (CESCs) in conjunction with NFV and Mobile Edge Computing (MEC) technologies, able to accommodate the requirements of different use cases associated to vertical industries. The management of the 5G ESSENCE infrastructure relies on the use of monitoring data and analytics as a fundamental component to characterise the behaviour of both the radio and cloud components of the network. In this context, the objective of this paper is to present the framework for telemetry and analytics that is being developed by the 5G ESSENCE project, identifying its main components in relation to the architecture considered by the project and illustrating its capabilities for supporting different optimisation use cases.

The paper is organised as follows. Section II presents the 5G ESSENCE architecture, introducing the main concepts considered in the project. Then, Section III describes in detail the components of the proposed monitoring and analytics framework. The capabilities of the framework are illustrated with two use cases. The first one, presented in Section IV, addresses a workload prediction mechanism as a tool for analytics-based optimisation of cloud resource allocation. Section V presents the second example, focused on the optimisation of the Radio Access Network through an energy saving function. Conclusions are summarised in Section VI.

II. 5G ESSENCE ARCHITECTURE

5G ESSENCE architecture intends to support the provision of SCaaS together with MEC capabilities in multi-tenant environments. The architecture allows multiple network operators (tenants) to provide services to their users through a set of CESCs deployed and owned by a third party (i.e., the CESC provider). In this way operators can extend the capacity of their own 5G Radio Access Network (RAN) in areas where the deployment of their own infrastructure could be expensive and/or inefficient, as it would be the case of e.g. highly dense areas where massive numbers of Small Cells would be needed to provide the expected services.

A CESC is a multi-operator and multi-RAT small cell (SC) that integrates a virtualized execution platform for executing novel applications and services inside the access infrastructure. It includes a SC Physical Network Function (PNF) unit supporting a subset of the SC functionalities and a micro server that supports the execution of Virtualized

Network Functions (VNFs) and provides the rest of the SC functionality (e.g. distributed Radio Resource Management [dRRM] and distributed Self-Organizing Network [dSON] functions) together with other added-value services (e.g. virtual Deep Packet Inspection [vDPI], low latency Machine-to-Machine applications).

The physical aggregation of a set of CESCs, denoted as a CESC cluster, allows the joint operation of the computational, storage and networking resources of the micro servers as a single Network Function Virtualised Infrastructure (NFVI), denoted as Light Data Centre (DC). To further extend the computational capabilities of this infrastructure, a centralized Main DC is also included, able to host more computationally intensive tasks and processes in order to have a global view of the underlying infrastructure. In addition to different network services composed of VNFs, the Main DC hosts a centralized Software Defined-Radio Access Network (cSD-RAN) controller that includes centralized Radio Resource Management (cRRM) and centralized Self-Organized Network (cSON) functions operating at multi-cell level.

The two-tier cloud architecture resulting from the combination of the Light DC and the Main DC, denoted jointly as Edge DC, provides a flexible NFVI platform for efficiently hosting MEC services and for virtualizing the SC functionalities. As for the technology to be used in the NFVI to realise VNFs, the default option is to use Virtual Machines (VM). But with the advent of new virtualization techniques, other options are being considered, including the use of containers and unikernels [6].

The main management component of the architecture is the so-called CESC Manager (CESCM). Firstly, it hosts the Element Management System (EMS), which provides a package of end-user functions for the management of both the PNFs and VNFs at the CESCs, and the Network Management System (NMS) for managing the whole set of CESCs deployed by an operator. Secondly, the CESCM also includes the Network Function Virtualisation Orchestrator (NFVO), in charge of creation and management of network services on the virtualised infrastructure. For that purpose, the NFVO abstracts the services in the virtual networking environment and, with the support of the VNF Manager (VNFM), it instantiates, updates, queries, scales and terminates the VNFs.

In 5G ESSENCE, OSM - Open Source MANO (Management and orchestration) - [7], which encompasses both the NFVO and VNFM, is used. OSM provides the required level of flexibility to implement MANO functionalities to the infrastructure and a high technical readiness level. Fig. 1 depicts the orchestration tool considered in 5G ESSENCE, together with the rest of components used for lifecycle management of network services. The next element to be found when moving one level down in the orchestration stack is the Virtualised Infrastructure Manager (VIM) – OpenStack is used in 5G ESSENCE -, which executes the directives coming from the orchestrator, over the virtual resources. From the network perspective, a Software Defined Network (SDN) controller works with the forwarding plane and adjusts network configuration to provide the connections between different VNFs or resources. In the case of 5G ESSENCE, OpenDaylight is the selected option for this controller due to the fact that it supports OpenFlow, but also other open SDN standards.

Finally, 5G ESSENCE encompasses a monitoring and analytics framework that captures and analyses relevant indicators of the network operation. This provides the CESCM with accurate knowledge models that characterise the behaviour of the network and its users in relation to the utilisation of both cloud and radio resources and therefore, it supports the orchestrator and the cSD-RAN elements when making decisions. This framework is presented in details in the next section.

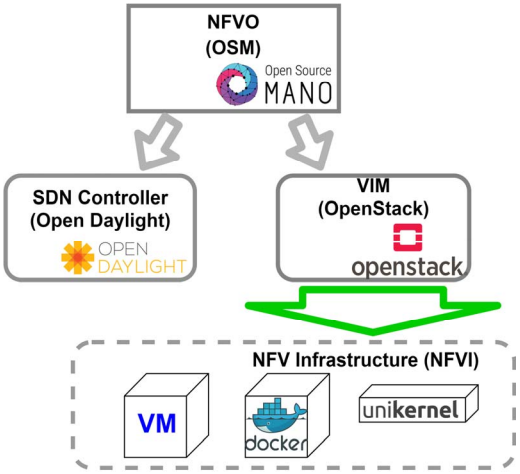

Fig. 1 Tools for orchestration in 5G ESSENCE

III. Monitoring and Analytics Framework

The 5G ESSENCE system presents a high degree of dynamicity, due to the constantly changing behaviour of services and workloads to be supported by the radio and cloud infrastructure. From this perspective it is required a proper monitoring system able to adapt to the different supported scenarios. It is very important to provide the orchestration framework with insights generated by an understanding of which physical and virtual resources are available on the infrastructure and their status over time. The data collected by the monitoring system is used for visualization purposes (for human consumption) and is also provided to a set of analytics techniques capable of extracting insights from the data and, via feedback loop, enabling the realisation of efficient resource allocation across the infrastructure.

One of the challenges for telemetry in 5G small cell deployments is that they are characterised by a high degree of distribution. For example, in the 5G ESSENCE architecture, the infrastructure functionalities are distributed across different nodes of the network going from a centralised location (the Main DC) to the very edge of the network (Light DC) where the small cells are deployed. Then, to achieve a full end-to-end view of both infrastructure and services, a shift in the design principles of telemetry is required, moving from a highly compartmentalized interpretation to one which utilizes metrics across all constituent elements. This requires the telemetry system to be able to instrument and monitor the different devices composing the overall infrastructure and to provide a unique and simple-to-access view of the system that can be exposed to both dashboards and analytical techniques.

Another challenge is related to the high complexity which characterises the overall telemetry system: the huge number of metrics and volume of data to be collected, processed and analysed increases the complexity of the

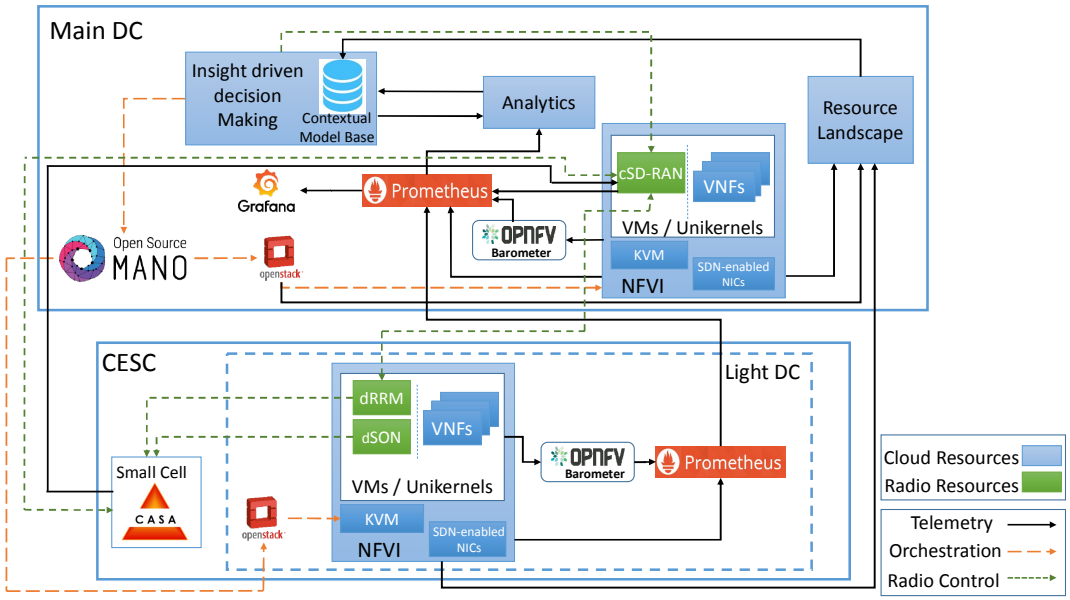

Fig. 2 Components of the 5G ESSENCE telemetry and analytics framework

decision making modules, slowing down the reaction time of the system and implicitly increasing service latency. The telemetry system is then required to be distributed in nature and to adapt to the constantly changing needs of the infrastructure. For this purpose, two key desired characteristics of the telemetry platform are: (i) the capability to generate aggregated and derived metrics and (ii) the capability to store and consequently access the data locally using a distributed monitoring approach.

Based on the above requirements, Fig. 2 depicts the main components of the 5G ESSENCE telemetry and analytics framework and their placement in relation to the Main DC and Light DC. The Analytics component is envisioned as a collection of Machine Learning and data analytics tools for the generation of models to be used for provisioning of insights to the orchestration system, i.e., OSM, and to the cSD-RAN controller of the small cells. The collection of these models is represented in the insight driven decision making box of Fig. 2.

The analytics component within the Main DC is fed by telemetry data. In this respect, different publicly available telemetry platforms have been analysed in relation to the 5G ESSENCE requirements. They include Graphite [8], Nagios [9], Prometheus [10], CollectD [11], OPNFV Barometer [12], Zabbix [13] and the OpenStack Monitoring Framework [14]. Among them Prometheus has been selected as the most suitable solution to monitor the two tier infrastructure for implementation purposes, since it provides functionalities to cover all the aforementioned requirements.

Prometheus monitoring tool provides a modular architecture that includes built-in and active scraping, storing, querying, graphing, and alerting based on time series data, with a large support of open source community. Those features provide a complete framework to handle different types of infrastructures separately and to analyse the end-to-end services in the architecture. Furthermore, an important ability of Prometheus allows defining alerts on those metrics, to be notified over multiple channels, in case of any discrepancy or event with the infrastructure. For this purpose, Prometheus connects with different exporters to collect the pertinent metrics of the service composition. Each exporter works as an independent service that remotely gathers information about the infrastructure, application logs or other interesting data to collect, accumulating this data and exposing it to Prometheus through HTTP. Moreover, alerts are defined by using Prometheus Alert Manager. It provides level of control and management workflow, allowing the platform to warn management entities, as endpoints, of problems identified in the infrastructure that can affect the QoS.

As shown in Fig. 2, in some cases the Prometheus platform will be coupled with the usage of CollectD plugins developed within the OPNFV Barometer project [12] in order to support collection of VM and network telemetry metrics (libvirt, Data Plane Development Kit [DPDK], etc.).

Prometheus queries can support basic logical and arithmetic operators, such as addition, subtraction, multiplication, division, modulo and exponentiation. It also supports processing of metrics using aggregation operators, such as the calculation of the sum, the minimum, the maximum, the average, the standard deviation, the variance and the quantile over dimension, as well as providing keywords for counting values and filter them (such as taking the smallest and largest k elements by sample value). This supports the capability to generate aggregate and derived metrics.

In addition to the default deployment consisting of a centralised single instance which is in charge of collecting all the data exposed by the overall infrastructure, Prometheus provides a functionality called Federation, which supports the distributed monitoring and storage of data. The hierarchical federation allows Prometheus to scale in environments with tens of data sources and potentially millions of resources. The federation topology looks like a tree, with higher-level Prometheus servers collecting aggregated time series data from a larger number of subordinated servers.

A fundamental aspect of the 5G ESSENCE monitoring and analytics framework is the capability of monitoring the RAN, as a difference from most of the previously mentioned telemetry platforms, which have typically been used for the collection of measurements related to Information Technology (IT) infrastructure components,

e.g., CPU (Central Processing Unit) usage, memory, operating systems, etc. In contrast, the monitoring of the RAN involves collecting radio interface-related measurements from the CESCs.

5G ESSENCE considers a multi-RAT CESC that supports Long Term Evolution (LTE), 5G New Radio (NR) and/or Wi-Fi technologies. In this respect, LTE small cells collect a number of measurements at the physical (PHY) layer (see [15]) and at the Layer 2 (see [16]). These measurements can be collected by the small cell or by the User Equipment (UE) that reports them to the small cell. Similarly, for 5G NR the list of PHY measurements is given in [17], while Wi-Fi measurements are described in [18].

Although the above measurements are available at the small cells, the capability of a telemetry platform like Prometheus to collect them depends on the actual configuration of measurements that each small cell exposes to an external system, typically through management interfaces defined between the small cell and the EMS and/or NMS. The interface between the small cells and their EMS is typically vendor specific, but there have been some efforts in defining open standard interfaces such as TR-196 [19] supported by multiple vendors. Similarly, 3GPP has also standardized different Performance Measurement (PM) metrics [20] and Key Performance Indicators combining these metrics [21] to be transferred from the small cells (or their EMS) to the NMS. These PM metrics are provided in the form of XML files following the format of [22] and produced according to a configured reporting interval. Each file can contain one or more granularity periods, which define the time across which measurements are collected and aggregated. In 5G ESSENCE, PM files are generated by the cSD-RAN controller. It exposes the relevant metrics to a Representational State Transfer (REST) Application Programming Interface (API) that generates a JavaScript Object Notation (JSON) service that is then translated by a specific exporter, which consumes and translates the data to be understandable by Prometheus.

Telemetry data obtained from the RAN and the subsequent analytics functions executed on this data can support the decisions made by different RRM/SON functions corresponding to the green boxes of Fig. 2. The dRRM component shown in Fig. 2 represents virtualized RRM functions running at the Light DC. They can be functions such as admission/congestion control, load balancing, etc. operating at the time scale of the data session duration or even longer. By exposing these functions outside of the physical small cell using proprietary interfaces within Medium Access Control (MAC) layer, it becomes possible to literally software-define some of these functions that are currently vendor locked. More fine grained RRM functions (e.g., scheduling) could be virtualised at the Light DC only if strict real-time requirements could be ensured through the interface between the SC PNF and the Light DC, for instance, via usage of shared-memory communications and the Femto Application Platform Interface (FAPI). A similar discussion applies for the dSON component of Fig. 2 that enables local optimisations of resource management policies (thresholds, parameter settings) at cell level.

The presence of dRRM and dSON components at the Light DC provides a natural extension of the key functionalities of the cSD-RAN controller, leading to a hierarchical implementation of RRM. This will benefit from decentralized decision making –higher level decisions can be undertaken at Main DC while local implementation of necessary optimisations most suitable in a given small cell context happens near to the cell. The distributed deployment of RRM and SON functionalities as virtual components to be deployed on the Light DC gives control to the NFVO over the resource allocation and management for those functionalities. In fact, the orchestrator will be able to determine optimal VNFs' location based on the current resource state (telemetry) and the optimisations of workloads (analytics) and use those for the ultimate placement of functions between the two DCs.

The monitoring framework of 5G ESSENCE is designed to also collect information about the cloud resources. First the framework requires supporting the identification of available physical resources, which will be allocated to VNF Components and then it requires monitoring the interdependency between virtual and physical resources. In the 5G ESSENCE monitoring framework, this functionality is covered by the Resource Landscaper shown in Fig. 2. The Landscaper comprises of a landscaper aggregator, collector-agents and collector components. The collector-agents run on physical compute nodes and are configured to use one or more types of collectors. The collectors gather different types of information that are passed to the collector-agents and then onto the landscaper aggregator that stores that information on a database.

Relationships among resources are expressed by representing the overall infrastructure as a graph, in terms of nodes and edges, where every node represents a resource and where links between resources represent various forms of relationships. Contextual information about every resource is expressed by means of attributes that include representation of the features of each resource, but also metadata that describe the type of node. The nodes are in fact of different categories, such as compute, network and storage, and of different types, such as CPU, core, Non Uniform Memory Access (NUMA) node, Peripheral Component Interconnect (PCI) Bridge, Network Interface Card (NIC), etc. Every type of nodes comes with specific attributes (e.g. the CPU is characterized by a given frequency, a NIC is characterized by a given throughput, etc.). Also every resource belongs to a logical layer, which can be physical, virtual or service. For instance CPU, NIC, disk are, example of physical nodes, VMs vNICs and virtual Networks are an example of virtual nodes and stacks are examples of service nodes. Prometheus collects telemetry information for those components, enabling the runtime decision making based on the state of cloud resources.

IV. Analytics-Based Optimisation of Cloud Resource Allocation

The operation of the orchestration, including the management of the initial service placement, scaling and migration, will be influenced by the information coming from the monitoring and analytics module, combined with the contextual information acquired over time. In this respect, this section presents a specific use case related to the design of the analytics to add the capability for the system to predict the behaviour of workloads and users and use that prediction for intelligent decision making.

We investigate the idea of integrating system telemetry ranging from standard resource usage statistics (CPU usage, memory consumption, etc.) to kernel and library calls of

applications into a machine learning (ML) model to approximate, at any point in time, the state of a system and allow us to solve tasks such as resource usage prediction and anomaly detection. To achieve this goal, we train recurrent neural networks such as Long Short-Term Memory (LSTM) networks [23] to learn a model of the monitored system.

To learn a model of a system that is good at predicting future resource usage, we need to collect data about the present. One obvious approach is to collect data about the resources that we want to predict, such as CPU and memory usage statistics. This follows the idea that, in many cases, the previous values of resource usage will have at least some influence on current resource usage. Such resource usage information is easily available; however, it is generally accounted for on a per-process basis. However, logical services often comprise of a number of processes running in sequence or parallel for which is therefore necessary to aggregate measurements to create a global view of the service, a task that Prometheus can assist with within the 5G ESSENCE architecture.

In order to prepare the data such as usage statistics as input for the ML models, we need to discretize it into time intervals. Fig. 3 depicts the overall architecture: we collect telemetry data (top left), as well as the application's system calls (bottom left), over a time period t. The variable number of calls is transformed into a fixed-size vector via representation learning as described in [24]. Within the scope of this work, a skip-gram model such as word2vec [25] to create a word embedding has shown good performance. The two vectors are combined and used as input for an LSTM that then predicts the resource usage at some future time period $t + i$.

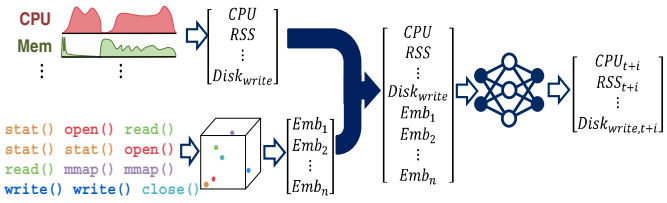

Fig. 3 Architecture of a system that takes telemetry and system call data to predict future resource usage

Fig. 4 shows initial results of the prediction algorithm. They were obtained by collecting system calls from various applications to create the system call corpus. We then collected the resource usage and system calls of a scientific computing toolchain that executed a number of bash and python scripts, which interleaved I/O-and CPU-heavy phases. Finally, we embedded the system calls and trained an LSTM with the data. The model is trained to minimize the RMSE of the CPU usage (as a value between 0 and 1) i seconds into the future. Results of Fig. 4 are obtained by varying both how far to predict into the future, and how much history to take into account for the prediction.

Results reflect that with the reduction of history taken into account for the LSTM, the prediction error increases as there is less information available for the model to base its prediction on. Besides, it would be expected that looking farther into the future would increase the prediction error (since uncertainty increases with the amount of time between the two points in time). However, this only holds true to a very limited degree: for the 1-second-of-history case, the error indeed increases, but only up to a prediction of seconds into the future, after which it levels out. For the case of 10 seconds of history, the accuracy stays roughly the same over all investigated prediction lengths.

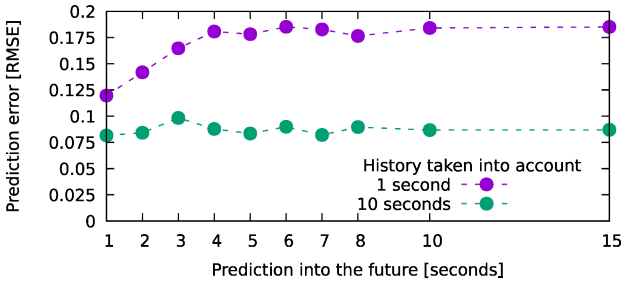

Fig. 4 Initial results of the prediction algorithm

V. Analytics-based Optimisation of the Radio Access Network

This section presents a use case to illustrate the introduction of analytics in the RAN. It focuses on the analysis of the cell-level time domain traffic pattern, which defines how the traffic of a small cell varies as a function of time. The traffic is tightly related with the environment where the cell is deployed and with the characteristics and profiles of the users served by the cell. Therefore, the detailed analysis of traffic patterns allows extracting valuable knowledge that can be used for making management decisions regarding the configuration of a cell.

A specific area of applicability of the knowledge extracted from cell-level time domain traffic patterns is the energy saving function, which intends to reduce the energy consumption in the RAN by switching off the cells that carry very little traffic at certain periods of the day (e.g. at night) and making the necessary adjustments in the neighbour cells so that the existing traffic can be served through some other cell. In this respect, we consider here an example of analytics-based approach to automate the identification of the small cells that are candidate to be switched off during certain periods of the day. The automation is based on a supervised classification process that allows incorporating human expert criteria.

Fig. 5 illustrates the considered approach. The cSD-RAN controller generates PM reports of the small cells in the form of XML files. Among different metrics, these PM reports include the number of active UEs in each cell [20]. The telemetry platform Prometheus would collect the results of the PM reports and, for each cell i, it would build a time series $\mathbf{X_i}$ with the time samples of the measured traffic (i.e. number of active UEs in the cell) at different time instants corresponding to the granularity periods of the measurements included in the PM files (e.g. 15 min). Besides, the time series corresponds to a given duration (e.g. 1 week in the example considered here).

Based on the collected data $\mathbf{X_i}$, the objective of the analytics is to perform a classification process of each cell to determine its class $C(\mathbf{X_i})$, which can be A (Candidate cell to be switched off) or B (Cell that cannot be switched off). Prior to that, and to reduce the dimension of $\mathbf{X_i}$, an initial processing of time series $\mathbf{X_i}$ would be carried out exploiting the aggregation capabilities of Prometheus. Specifically, this processing computes a vector $\mathbf{F(X_i)}$ with the average normalized traffic of the cell during the nights (i.e. from 0h to 8h), mornings (i.e. from 8h to 16h) and afternoons (i.e. from 16h to 24h) for each day of the week. The vector $\mathbf{F(X_i)}$ is provided to the main analytics module to perform the

classification process, i.e. the association between the input $\mathbf{F(X_i)}$ and the class $C(\mathbf{X_i})$ of each cell.

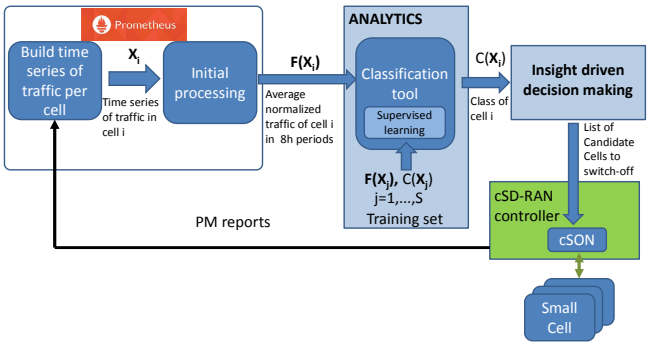

Fig. 5 Telemetry and analytics stages for performing the cell classification for energy saving

The internal structure of the classifier depends on the specific classification tool being used (e.g. Support Vector Machines, Neural Network, Naive Bayes, Decision Tree, etc.) and its settings are automatically configured through a supervised learning process executed during an initial training stage. This process uses as input S time series $\mathbf{X_j}$ of some cells whose associated classes $C(\mathbf{X_j})$ are pre-defined by an expert. In this way, the training set will be composed by the tuples $(\mathbf{F(X_j)}, C(\mathbf{X_j}))$ j=1,...,S. The supervised learning process analyses this training set to determine the appropriate configuration of the classification tool. This will allow automating the classification process for other cells not included in the training set. The RapidMiner tool [26] is considered here for implementing the classification process.

To illustrate the abovementioned process, Fig. 6 depicts the time domain traffic pattern $\mathbf{X_i}$ of two cells, one classified as A and another as B, which has been taken from [27] using a Support Vector Machine classifier. It is observed that the class A cell exhibits relatively long periods at night serving no traffic at all, while the class B cell has traffic during all the time periods.

The outcome of the analytics is the list of small cells with their associated class. Then, following the general monitoring and analytics framework of Fig. 2, this information will be stored in the insight driven decision making and will be delivered to the energy saving cSON function at the cSD-RAN controller (see Fig. 5). This cSON function would be in charge of performing the final decision of switching off the cells and reconfiguring their neighbours appropriately.

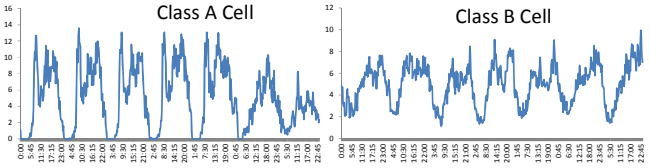

Fig. 6 Example results of the traffic evolution during one week for one cell classified as A and another one classified as B.

VI. Conclusions

This paper has presented the monitoring and analytics framework that is being developed by the 5G ESSENCE project to support the management of small cells as a service provision through multi-tenant cloud-enabled small cells. The framework is based on an analytics component that incorporates different statistical and machine learning tools for supporting the decision making of both the orchestration and the RAN management and control. Analytics are fed by telemetry data collected from the NFVI, the VNFs and from the RAN.

Two different use cases have been provided to illustrate the proposed framework. The first one addresses the use of analytics for developing a prediction algorithm of resource usage. The second one presents a classification mechanism to identify the small cells that can be switched off for energy saving purposes at certain periods of the day.

Acknowledgement

This work has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 761592 (5G ESSENCE project).

References

[1] NGMN Alliance, "5G White Paper", February, 2015.
[2] ETSI GS NFV-MAN 001 v1.1.1, "Network Functions Virtualisation (NFV); Management and Orchestration", December, 2014.
[3] C-L I., Y. Liu, S. Han, S. Wang, G. Liu, "On Big Data Analytics for Greener and Softer RAN", IEEE Access, August, 2015.
[4] M. J. McGrath, V. Bayon-Molino, "Evolving end to end telemetry systems to meet the challenges of softwarized environments", IEEE Softwarization, May, 2017.
[5] http://www.5g-essence-h2020.eu/
[6] A. Madhavapeddy, D. J. Scott, "Unikernels: Rise of the Virtual Library Operating System", Magazine Queue-Distributed Computing, Vol. 11, No. 11, November, 2013.
[7] https://osm.etsi.org/
[8] https://graphiteapp.org/
[9] https://www.nagios.org/
[10] https://prometheus.io/
[11] https://collectd.org/
[12] https://wiki.opnfv.org/display/fastpath/Barometer+Home
[13] https://www.zabbix.com/
[14] https://wiki.openstack.org/wiki/Telemetry
[15] 3GPP TS 36.214 v15.1.0, "Evolved Universal Terrestrial Radio Access (E-UTRA); Physical layer; Measurements (Release 15)", March, 2018.
[16] 3GPP TS 36.314 v15.0.0, "Evolved Universal Terrestrial Radio Access (E-UTRA); Layer 2 - Measurements (Release 15)", March, 2018.
[17] 3GPP TS 38.215 v15.1.0, "NR; Physical layer measurements (Release 15)", March, 2018.
[18] IEEE Std 802.11-2016, "Part 11: Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specifications", December, 2016.
[19] TR-196 "FAPService: 2.0 Femto Access Point Service Data Model", Broadband forum, https://cwmp-data-models.broadband-forum.org/tr-196-2-0-0.html, Accessed June 2018.
[20] 3GPP TS 32.425 v15.0.0, "Performance Management (PM); Performance measurements Evolved Universal Terrestrial Radio Access Network (E-UTRAN) (Release 15)", March, 2018
[21] 3GPP TS 32.450 v14.0.0, "Key Performance Indicators (KPI) for Evolved Universal Terrestrial Radio Access Network (E-UTRAN): Definitions (Release 14)", April, 2017.
[22] 3GPP TS 32.435 v14.0.0, "Performance measurement; eXtensible Markup Language (XML) file format definition (Release 14)", April, 2017.
[23] S. Hochreiter and J. Schmidhuber, "Long Short-Term Memory," Neural Computation, vol. 9, no. 8, pp. 1735--1780
[24] F. Schmidt, M. Niepert, F. Huici, "Representation Learning for Resource Usage Prediction", SysML Conference, Stanford, CA, USA, February 2018
[25] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, J. Dean, "Distributed representations of words and phrases and their compositionality", NIPS '13, Lake Tahoe, NV, USA, December 2013
[26] https://rapidminer.com/
[27] J. Pérez-Romero, J. Sánchez-González, O. Sallent, R. Agustí, "On Learning and Exploiting Time Domain Traffic Patterns in Cellular Radio Access Networks" 12th International Conference on Machine Learning and Data Mining (MLDM), New York, USA, July, 2016.