Received January 10, 2019, accepted February 10, 2019, date of publication March 5, 2019, date of current version April 9, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2902984

End-to-End Data Analytics Framework for 5G Architecture

EMMANOUIL PATEROMICHELAKIS[1], FABRIZIO MOGGIO[2], CHRISTIAN MANNWEILER[3], PAUL ARNOLD[4], MEHRDAD SHARIAT[5], MICHAEL EINHAUS[6], QING WEI[1], ÖMER BULAKCI[1], AND ANTONIO DE DOMENICO[7]

[1]Huawei GRC, 80992 Munich, Germany
[2]Telecom Italia, 20123 Milan, Italy
[3]Nokia Bell Labs, 81541 Munich, Germany
[4]Deutsche Telekom, 64295 Darmstadt, Germany
[5]Samsung Electronics R&D Institute, Staines TW18 4QE, U.K.
[6]Department of Communications Engineering, Leipzig University of Telecommunications, 04277 Leipzig, Germany
[7]CEA LETI, 38054 Grenoble, France

Corresponding author: Emmanouil Pateromichelakis (emmanouil.pateromichelakis@huawei.com)

This work was supported by the H2020-ICT-2016-2017/H2020-ICT-2016-2 through the project 5G-Monarch under Agreement 761445.

ABSTRACT Data analytics can be seen as a powerful tool for the fifth-generation (5G) communication system to enable the transformation of the envisioned challenging 5G features into a reality. In the current 5G architecture, some first features toward this direction have been adopted by introducing new functions in core and management domains that can either run analytics on collected communication-related data or can enhance the already supported network functions with statistics collection and prediction capabilities. However, possible further enhancements on 5G architecture may be required, which strongly depend on the requirements as set by vertical customers and the network capabilities as offered by the operator. In addition, the architecture needs to be flexible in order to deal with network changes and service adaptations as requested by verticals. This paper explicitly describes the requirements for deploying data analytics in a 5G system and subsequently presents the current status of standardization activities. The main contribution of this paper is the investigation and design of an integrated data analytics framework as a key enabling technology for the service-based architectures (SBAs). This framework introduces new functional entities for application-level, data network, and access-related analytics to be integrated into the already existing analytics functionalities and examines their interactions in a service-oriented manner. Finally, to demonstrate predictive radio resource management, we showcase a particular implementation for application and radio access network analytics, based on a novel database for collecting and analyzing radio measurements.

INDEX TERMS 5G, architecture, data analytics, network slicing.

I. INTRODUCTION

The fifth generation (5G) mobile communications system is characterized by a wide-range of services grouped under three generic service types, namely, enhanced mobile broadband (eMBB), massive machine-type communications (mMTC), and ultra-reliable and low-latency communications (URLLC). The network slicing concept is introduced in 5G to address the various requirements from multiple vertical industries assuming a shared physical network infrastructure. A network slice can be customized according to the needs of vertical industries and services to be supported. Network slicing is a key pillar in 5G networks. The end-to-end (E2E) nature imposes domain-specific requirements that will span over multiple technical domains, i.e., radio access network (RAN), transport network, and core network (CN). In addition, 5G shall be supported by a management and orchestration (M&O) layer in order to meet defined service-level agreements (SLAs) for network slices of different nature. In 3rd generation partnership project (3GPP), four standard slice/service types (SSTs) have been introduced, namely, eMBB, mMTC, URLLC, and vehicle-to-everything (V2X) SSTs, which aim to provide differentiated handling for different slice specific data flows. An overview of network slicing enabler and interface enhancements in the 3GPP architecture is provided in [1], and the notion of slicing in emerging 5G network is presented in [2].

The associate editor coordinating the review of this manuscript and approving it for publication was Kostas Psannis.

To jointly accommodate challenging performance and operation requirements of multiple slices, the network functions (NFs) and resources are envisioned to be stretched, as the complexity and signaling requirements may significantly increase to optimize the network in a slice-tailored manner. One particular enabler for relaxing the signaling/ complexity requirements and at the same time to enhance the functional and resource utilization is the notion of *Data Analytics*. The derived statistics of network resource usage can serve as additional input for improving the 5G system (5GS) management, and for some use cases, the service operation efficiency, allowing the 5GS to re-act fast to dynamic network or service adaptations.

This article aims to investigate how data analytics can be used to enhance the 5GS performance and its impact on the architecture. In this context, we initially provide an overview of the 5G architecture with particular focus on data analytics, and explicitly describe the requirements for data analytics in 5G, assuming different optimization objectives set by the network operator or vertical customers. Given the requirements for network and service optimization, which can have multiple flavors subject to the service/slice requirements and the network capabilities, we further propose an integrated data analytics framework.

This framework uses as basis the current 5G architecture that supports network data analytics functionality (NWDAF) and management data analytics functionality (MDAF). On top of that, to employ analytics for real-time operations and allow vertical customers to have a tight control on some network operations, we propose additional functionalities for RAN-centric, data network-centric, and application-level analytics. For these functionalities, their integration in current architecture is examined, and subsequently the required extensions in standardized functions and interfaces are investigated.

Finally, a case study that requires data analytics at RAN to perform real-time resource management is evaluated, using application-level analytics on the user mobility, based on real measurements collected in a database residing at RAN. This trend of per User Equipment (UE) radio conditions given the expected routes are analyzed and output to the scheduler at the base station, in order to can dynamically adjust the Quality of Service (QoS) attributes or pro-actively adapt the resource allocation for a given time window. The use of analytics for the resource optimization can have significant impact in V2X safety-related and low-latency services, where the complexity and signaling costs strongly affect the performance in dense RAN deployments.

The remainder of this article is organized as follows. Initially, Section II, discusses a classification of data analytics based on the various factors, like the type, granularity, and objective. Following that, in Section III, we present the requirements for data analytics in 5G networks and classify them based on their impact on the different parts of 5GS and involved stakeholders. In Section IV, the status of analytics in 5G is presented focusing on the 3GPP system, use cases, and architecture. This extensive 3GPP review describes the current support of analytics from CN, RAN, and M&O perspective. Furthermore, in Section V, an E2E framework of data analytics functionality support is proposed, that allows to customize the analytics functions in 5G system. In this context, an exemplary implementation of the integrated analytics architecture is proposed in Section VI to capture how some critical services can be benefited by RAN-centric data analytics in terms of radio resource optimization. Finally, in Section VII, some key open research directions are discussed, to capture challenges that require further study in both academia and standardization.

II. DATA ANALYTICS CHARACTERIZATION

Initially, as an attempt for classifying analytics to different blocks, the possible attributes of the data analytics functions are discussed, focusing on what parameters are expected to be monitored/predicted, what type of analytics are required and what the granularity/frequency of operation is.

A. MONITORED PARAMETERS

Firstly, we decompose the prediction/analytics functionalities in different levels, based on the predicted or expected parameter. This can involve a UE session, or the resource load/situation in a particular domain, or the application/ service operation.

- **UE/Session-related parameters**: These parameters may include the prediction of the UE context/behavior to enable the network to better provision the resources. One example can be the mobility of the user or group of users, which can be used for handover management, or the prediction of interference that the UE will suffer/cause in a particular area. One further example is the prediction of QoS for one or more UEs in a given area.
- **Network-related parameters:** Here, these parameters can be grouped based on the domain they apply. In **RAN**, parameters can include the ones regarding the radio resource conditions and availability (e.g., average channel quality, load, and interference) as well as the traffic (e.g., user density) and other factors in real-time or non-real time. In **transport/backhaul**, the parameters that can be estimated concern resource conditions, backhaul/ fronthaul (BH/FH) type, topology, availability, dynamicity, etc. Finally, for **CN,** some parameters that can be monitored can be about the processing load and availability of CN functions.
- **Service-related parameters:** This category includes the analytics which can be performed at the application domain (e.g., at terminal or at the application function) and may be used by the 5G network to improve the service operation. One example which is specific for V2X slicing case, is the prediction of UE trajectory/route, traffic conditions, or expected Level of Automation (LoA) for a particular area.

- **Management-related parameters:** This category includes Performance Management (PM) and Fault Management (FM) analytics as introduced in 3GPP SA5. This parameter may take into account the current slice/subnet performance and statistics on, e.g., radio failures and will provide analytics to MDAF.
- **Cloud-related parameters:** This includes the cloud processing parameters, e.g., the load and availability of computational resources, which may have impact on the decision for virtualization of NFs to cloud platforms. Here to mention that, in a distributed cloud-based architecture, the above categories of parameters may be deployed on demand in edge or core cloud platforms. Given the tight latency and reliability requirements of some virtualized NFs (e.g., in RAN domain), performing analytics on the estimated computational resource load/conditions is of key importance for performing actions, like offloading processing load to other cloud processing units.

B. GRANULARITY OF ANALYTICS

1) REAL-TIME

The analytics can be performed in real-time operations (e.g., channel prediction in ms time scale); however, this is more challenging task due to the fact that additional processing might be required and the overhead may affect performance.

2) NEAR-REAL TIME/NON-REAL TIME

In this case, the analytics, is performed in sec/min/hour time scale and may apply to certain types of prediction (e.g., load distribution in a geographical area). In O-RAN [3], near-real time operations have been defined to capture operations like QoS management, traffic steering, and mobility management which may be semi-dynamic (e.g., hundreds of ms timescale).

3) ON DEMAND

This can apply to both real-time and non-real time analytics, and is the case when the vertical or the operator requires to enable this feature as a service, for a certain area or time window to meet the requirements of a network slice in a given area.

C. TYPE OF ANALYTICS

There are different types of analytics that can be useful for the network according to the Gartner's Graph on stages of data analysis [4]:

- Descriptive Analytics – Explaining what is happening now based on incoming data.
- Diagnostic Analytics – Examining past performance to determine what happened and why.
- Predictive Analytics – An analysis of likely scenarios of what might happen.
- Prescriptive Analytics – This type of analysis reveals what actions should be taken.

III. 5G REQUIREMENTS FOR END-TO-END ANALYTICS

Data analytics can be used to serve different purposes, depending on the time granularity of the derived statistics and the defined parameters it may support. This section aims to decompose the requirements for prediction functionalities in the 5GS based on different optimization objectives or expected benefits. This analysis consider different types of prediction models over different domains, e.g., CP, management plane, and service plane. The following subsections explicitly describe the key requirements for employing data analytics in the 5GS.

A. ANALYTICS FOR SERVICE ENHANCEMENT

3GPP SA1 has provided the 5G requirement (SMARTER) specification [5], where the key vertical use cases include V2X communication and vertical industry automation. For industrial automation, [6] has studied communications for factories of the future in 3GPP Release 16 including application areas and mapped applications (e.g., motion control, massive wireless sensor networks, augmented reality, process automation, connectivity for the factory floor, and inbound logistics for manufacturing). For some use cases, data analytics can be useful for ensuring network availability or for providing predictive maintenance features. Furthermore, for enhanced 5G-V2X scenarios such as extended sensor sharing and automated cooperative driving [7], 3GPP SA1 introduced the notion of network prediction, by enabling 5GS to notify a V2X application that the QoS of a UE's ongoing communication might need to be downgraded, e.g., due to predicted bad network conditions, change of radio technology, and radio congestion.

In such scenarios, the application QoS requirement and the network QoS offering may need to be negotiated so as to ensure that the network is able to offer certain service with given QoS to the end user. However, the interaction between network and application domain may not be as flexible and dynamic as required by the vertical to prevent performance degradation and service disruption, especially for critical services. In this direction, data analytics can be used to fulfill two different objectives:

- The 3rd party provides analytics on application session-related information per user or group of users, e.g., using authorized Application Functions (AFs) to the operator to support the network QoS and resource provisioning. This case is beneficial for some verticals, like V2X, where the application server has more control of some application-related functionalities, e.g., vehicle group formations, accurate positioning, etc. Such application-driven analytics may be in 3GPP scope for V2X and Industrial IoT (IIoT), and in particular within the 3GPP SA6 scope [8], which defines the application-layer support for V2X and future factories.
- The network provides network analytics w.r.t. future resources and QoS offering based on statistics and prediction algorithms residing at the network. The application server may subscribe or request to monitor the

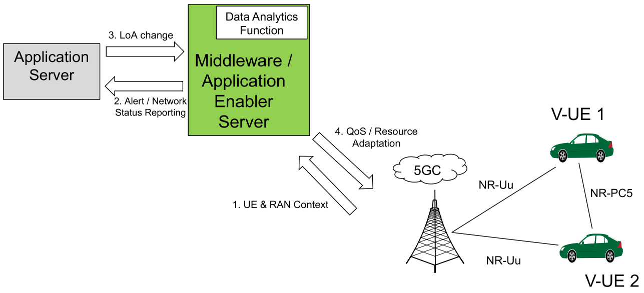


FIGURE 1. Application to network interaction using network data analytics at a middleware.

future network behavior to be able to pro-actively change the service offered to the end users and prevent loss or degradation of the service. A key example is the automotive vertical industry, which may require 5G network-assisted V2X communications for both safety and non-safety related services with diverse key performance indicators (KPIs). For enabling network analytics to improve V2X service/slice operations, 5GS reports predicted network changes (e.g., expected QoS degradation) to allow for adapting the level of automations at the vehicle side (e.g., by reducing the automation level to ensure slice/service continuity).

For both application and network driven analytics, as shown in Figure 1, one possible enabler for allowing the dynamic interaction between application and network could be a "middleware" functionality which can be a new layer between these domains, for translating the control and application requirements dynamically. Such functionality may reside in network or application domain and is under study in 3GPP SA6 (named as V2X Application Enabler Server [8]). This would be beneficial for V2X scenarios where the translation is essential for group communications and non-roaming multi-operator interactions. This functionality is also in-line with virtual RAN deployments (e.g., O-RAN architecture [3]), and could possibly reside as a 3rd party (e.g., original equipment manufacturer, OEM) application at a RAN Intelligent Controller. This allows the vertical customer to collect data analytics from the network side and trigger application-level adaptations or network-related QoS/ resource adaptations for single or groups of UEs.

B. ANALYTICS FOR 3GPP NETWORK ENHANCEMENT

5G has introduced network slicing support to allow for dedicated network configuration and optimization for individual scenario and services. However, due to the heterogeneous network (HetNet) deployment, the variation of network conditions as well as the changing of the traffic demand at different locations and times, the service assurance for each network slice may require complex network operation and management. This calls for the great simplification in network operation, e.g., via autonomy in network operation/control. Data analytics is one key enabler to support network automation. In the following sub-sections, we discuss the requirements for data analytics in 3GPP domains (RAN and CN domains), so as to enhance the network operation in 5GS.

C. RADIO ACCESS NETWORK DOMAIN

With its inherent characteristics, namely scarce resources (e.g., spectrum and computational resources), dynamic wireless channel conditions (e.g., fading and user mobility), necessity for low complexity in RAN deployments, along with the wide-spread utilization of higher frequency bands that are even more susceptible to radio conditions, the RAN may greatly benefit from data analytics in the 5G era. In addition, the RAN can be shared by a multitude of network slices, where the essential slicing objectives, such as slice isolation, service-level agreement (SLA) guarantee, and service continuity, shall be fulfilled [9]. To this end, data analytics can be utilized in the following use cases.

1) INTERFERENCE HANDLING

Data analytics, in particular diagnostic and prescriptive analytics can be utilized to enhance the performance of interference mitigation techniques. This can include selection of uplink (UL) power control parameters (e.g., fractional vs. full-compensation power control) and optimal number of blank sub-frames in case of time-domain interference coordination. Predictive analytics may support decisions in terms of the configuration of the initial parameters settings, e.g., considering a probable load increase due to group mobility. The utilization of data analytics can thus improve the resource utilization efficiency and reduce the need for frequent parameter adjustments.

2) PREDICTION OF CHANNEL QUALITY

Wireless channels are inherently dynamic in nature, which can be impaired by short-term and long-term fading. Especially, in case of mm-Wave communications the need for directional transmission introduces additional challenges where transmission collisions, i.e., experiencing high interference due to concurrent mm-Wave transmissions, can easily result in radio link failures (RLFs) [10]. Therefore, predictive analytics can be employed to decide on the active mm-Wave links such that the transmission collisions are minimized. To this aim, context information, e.g., geographical locations of the mm-Wave access points, antenna configurations, and power control parameter settings, can be utilized.

3) CONTROL OF DYNAMIC RADIO TOPOLOGY

One of the emerging 5G concepts is the use of unplanned dynamic small cells [10], where, for instance, a relaying functionality can be integrated into vehicles in the form of vehicular nomadic nodes (VNNs). VNN can be activated on-demand to adapt to the spatially and temporally changing traffic demands. Such deployment is dynamic in nature and requires tight network control, where the activation or deactivation of VNNs can depend on the channel conditions, vehicle battery, and the demand in a target service region.

Descriptive analytics can provide the needed context information to a dynamic RAN control unit, which determines the active VNNs. Moreover, predictive analytics can process the parking statistics in a certain region, which can be in turn used to predict the parking duration that sets an upper limit for the VNN availability.

4) ENERGY EFFICIENCY IMPROVEMENTS

Turning off the under-loaded cells fully or partially used jointly with load balancing is an effective way for improving the network energy efficiency. Predictive analytics can be utilized to determine the expected load of a cell or cell group such that the optimal trade-off between resultant energy efficiency gains vs. user performance reduction due to higher load of the active cells can be reached. The predictive analytics can, for instance, factor in the load history and activities in the neighbor cells.

5) MULTI-SLICE RESOURCE MANAGEMENT ENHANCEMENTS

As mentioned above, it is envisioned that RAN will be shared by a multitude of network slices with possibly diverse performance requirements, business-driven additional requirements, and slice-specific protocol configurations. This requires a common multi-slice resource management coupled with a tight performance monitoring to fulfill SLAs of the network slices, while ensuring resource isolation between slices. The resource isolation can be in time, frequency, code, and computational domains. Besides, slices with mission-critical services impose further constraints on the SLA fulfillment. On this basis, all types of the aforementioned data analytics shall be utilized. Namely, descriptive analytics can be part of the SLA monitoring, where different performance thresholds can be defined to avoid an SLA violation. Diagnostic analytics on the SLA violations can be used as input to prescriptive analytics to adjust the performance thresholds for violation prevention. In addition, predictive analytics can provide insights into slice load variations over time and space exploiting, e.g., the history information, which may be used for semi-static resource allocation and cell range extension parameters.

D. CORE NETWORK DOMAIN

In the fourth generation (4G) era, several descriptive and diagnostic data analytics are generated for NFs like the mobility management entity (MME) for mobility management, e.g., for deciding on the optimized paging area for selecting of an appropriate serving gateway (S-GW).

When it comes to 5G, there is significant room to put predictive and perspective analytics in usage as they enable an operator to predict an event (e.g., network overload and an upcoming outage or failure) earlier ahead to adopt suitable preemptive actions to ensure smooth network operation.

The current usage of data analytics is limited to the individual network function/entity at intra slice-level. However, E2E service assurance requires joint consideration on intra and inter-slice coordination of CP and management plane information, as well as the feedback from the application layer (e.g., 3rd party or PLMN-owned AFs authorized to closely interact with the 5GS). Based on these observations, a data analytics module should jointly consider the data from different NFs and different layers in a cross-slice manner.

3GPP currently standardizes what is known as Service-Based Architecture (SBA) [11] for the 5G Core as part of the System Architecture for 5GS in Release 15. The SBA defines different NFs and associated services can directly communicate with each other as originator or consumer of a service via a common bus known as Service-based Interface (SBI). For instance, a V2X AF may request or subscribe to the network for V2X-specific analytics information related to the expected network performance. Upon the change of expected network performance, the V2X AF may request devices to change from autonomous to human driving. Other examples include Network Slice Selection Function (NSSF) and Policy Control Function (PCF). The PCF may use per slice data analytics in its policy decisions. NSSF may use the load level analytic information for slice selection. Therefore, a new data analytics module is required to be deployed in 5G Core, which should be capable of providing per slice data analytics to multiple domains (e.g., other NFs, AFs, and operation, administration, maintenance - OAM) based on the on-demand slice-tailored data analytics requirements. Another key function envisioned within SBA is the NWDAF, enabling NFs to access to the operator-driven analytics for different purposes, including intelligent slice selection and control. More information on the evolution of network analytics in 3GPP will be covered as part of Section IV.

E. ANALYTICS FOR NETWORK MANAGEMENT ENHANCEMENTS

5G mobile networks are envisioned to be heterogeneous integrating all previous radio access network generations, while at the same time introducing new technologies such as New Radio and millimeter Wave (mm-Wave). 5G offers a variety of service assurance even when the bandwidth requirements and selected backhaul paths are the same, due to factors like the radio conditions, different access technologies, backhaul limitations, and connectivity options available in different areas.

A network slice may be requested to serve a greater geographic area. In that case, different type of equipment and technologies across the RAN, CN, and transport layer network, e.g., BH/FH, with diverse characteristics may be required. To assure a unified – homogeneous service provision across a network slice within a geographical area is a challenging problem and the solution is not straightforward, since this depends on a number of parameters including environmental, network technologies, load/users, etc. In addition, it is not obvious which resources to allocate and/or combine together in order to assure a stable desired SLA. The conventional ways to provide a unified – homogeneous service is resource overprovisioning, which

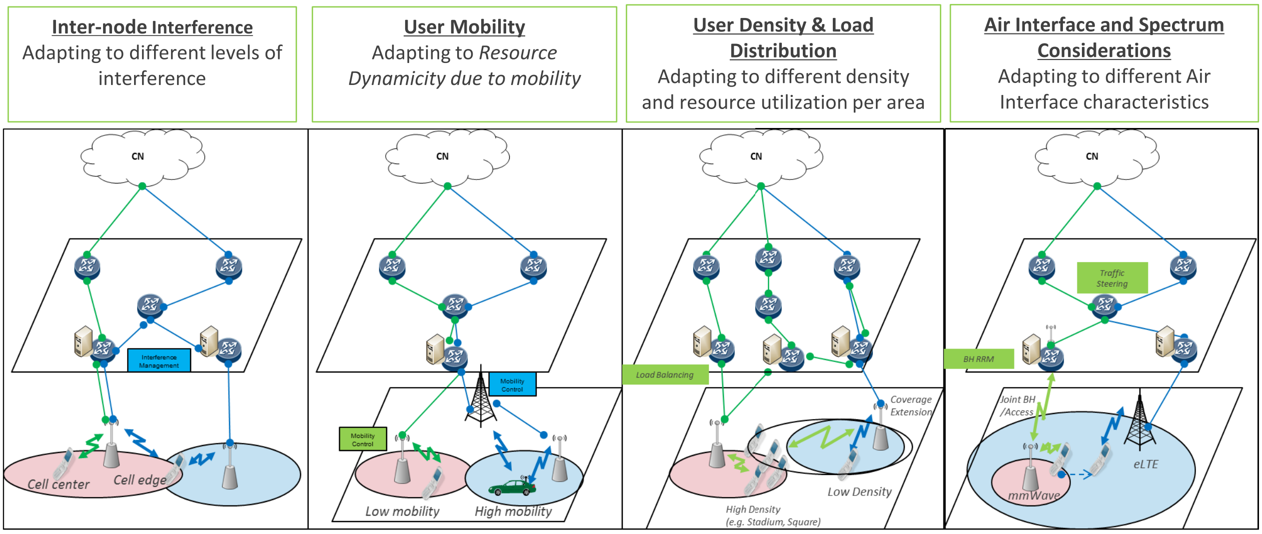

Inter-node Interference
Adapting to different levels of interference
User Mobility
Adapting to Resource Dynamicity due to mobility
User Density & Load Distribution
Adapting to different density and resource utilization per area
Air Interface and Spectrum Considerations
Adapting to different Air Interface characteristics
CN
Interference Management
Cell center
Cell edge
Mobility Control
Mobility Control
Low mobility
High mobility
Load Balancing
Coverage Extension
Low Density
High Density
(e.g. Stadium, Square)
Traffic Steering
BH RRM
Joint BH /Access
eLTE
mmWave

FIGURE 2. Data analytics use cases for network slice optimization.

increases the network costs. Another alternative is to quantify the E2E SLA considering the perceived user experience and to allocate the appropriate resource to ensure a stable SLA, requiring multiple domains to be synchronized and real-time information (per ms) to be acquired for the E2E scheduling decisions. This will provide very high signaling and complexity for enabling the network to synchronize different domains and allocate resources dynamically.

To this end, the use of data analytics and especially predictive analytics, for enabling the slice configuration and optimization can be beneficial for meeting the slice dynamicity and scalability requirements. However, the level and means of prediction can vary significantly based on the slice requirements (e.g., KPI) as well as the environment. Figure 2 shows different use cases where data analytics enables to adapt the slice configuration to interference, mobility, load, and air interface, in order to maintain the required level of quality of experience (QoE).

The slice requirements, along with the requirement of increasing the flexibility of the network to ensure homogeneous SLA across the slice coverage area, may present management and operational challenges and complexities when it comes to slice configuration and optimization. Therefore, the design of analytics should follow the below principles in the slice management domain:

- Network slice management shall be driven by complex data that are the outcome of aggregation / elaboration of signals coming from multiple network resources or slice subnets. Hence, the MDAF should be responsible for providing these processed data.
- Management data analytics operates at NF level, at network slice subnet level, and network slice level:
    1. Analytics at a NF level requires the collection of NF's load related performance data, e.g., resource usage status of the NF. This analysis could recommend appropriate configuration and lifecycle management actions, e.g., scaling of resources, admission control, load balancing of traffic.
    2. Analytics at network slice subnet level, shall provide information for closed-loop management of the subnet and information for the overall network slice management. The analytics service may further classify or shape the data in different useful categories and analyze them for different network slice subnet management, needs, e.g., scaling and admission control of the constituent NFs.
    3. Analytics at network slice level shall consume the analytics services exposed at slice subnet level to manage and orchestrate the slice life cycle in real time providing assurance management to the different communication services that are leveraging on the same shared network slice.
- The management data analytic service (MDAS) should utilize the network management data collected from the network (including e.g., service, slicing and/or NFs related data) and make the corresponding analytics based on the collected information to improve networks slice configuration and optimization. For example, the information provided by performance MDASs can be used to optimize network performance, and the information provided by fault MDASs can be used to predict and prevent failures at network slice level.
- Network slice configuration and optimization have to deal with the complexity of the management of shared resources and has to fit the different requirements and optimization needs coming from all the communication services that the slice has to support. In this regard, the management system may take into account what is happening in the network using real time data

analytics to decide the optimized configuration parameters which are used to create a new slice or to maintain a deployed one.

As mentioned above, the M&O layer has to deal with real time network allocation requests and real time assurance management for deployed network slices. Automation and self-organizing networks (SON) techniques (e.g., advanced algorithms for self-configuration, RAN network optimization, and self-healing) are the key concepts to effectively achieve such a goal. To further analyze how data analytics can be applied to optimize the access network, the 3GPP has defined two novel study items in SA5 and RAN3 workgroups respectively, namely "Study on SON for 5G" [12] and "RAN-Centric Data Collection and Utilization for long-term evolution (LTE) and NR" [13] for Release 16. The use cases (e.g., optimization of capacity/coverage, mobility and load sharing/balancing, energy saving, and minimization of drive tests) were primarily inherited from legacy systems. The outcome of the study items should include the identification of requirements and new signaling and interfaces required for SON enhancements for different 5G use cases. In particular, possible benefits and feasibility of introducing a logical RAN entity/function for data collection/utilization will be investigated in 3GPP RAN3, which could be interpreted as control or management functionality.

IV. 5G ARCHITECTURE AND USE CASES

In this section, the current 3GPP architecture and use cases with respect to data analytics are discussed. Specifically, the following sub-sections provide the overview of the work items w.r.t use cases as provided in 3GPP work groups to support data analytics in RAN, 5G CN, and M&O.

A. 5G RAN ARCHITECTURE

5G Phase 1 has been finalized, and the first features of RAN have been specified mainly for the eMBB scenarios, focusing on (non-standalone, standalone) architecture and protocol aspects [14]. One of the key design considerations for RAN is the flexibility of RAN deployments, where the 5G RAN access nodes (aka gNB) can be split horizontally (Central Unit (CU) – Distributed Unit (DU)) or vertically (CP – User Plane (UP)) to allow for flexible and service-tailored virtualization of RAN functionalities in different nodes, cf. Figure 3.

Based on the concept of virtualization, individual NFs within the RAN can be moved to telco cloud environments where computational resources are dynamically shared in order to increase energy efficiency and flexibility. The 5G RAN architecture will thereby most likely utilize both distributed cloud environment consisting of centralized clouds and edge clouds. The latter are also known as mobile edge clouds in the context of mobile radio networks.

The RAN protocol stack might be deployed virtualized in a telco cloud environment (cloud RAN). Even different NFs, such as the mobility management, resource scheduling, etc. for special service types, such as V2X, or a different application might run and share with each other the same hardware resources. The gNB within the 5G RAN can consist of a CU and one or more DUs, as specified in [15].

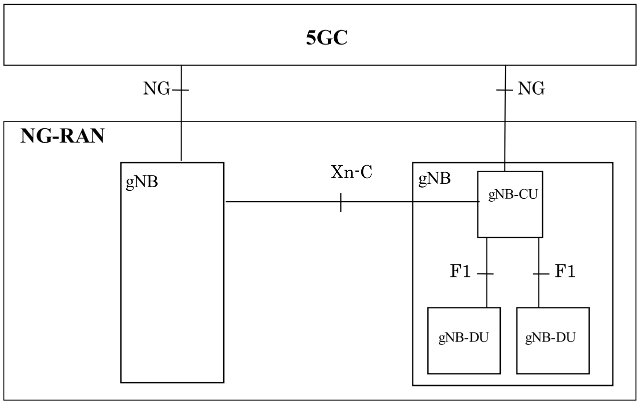


FIGURE 3. Illustration of 5G RAN Architecture connecting to 5G CN, aka 5GC, over NG interface [15].

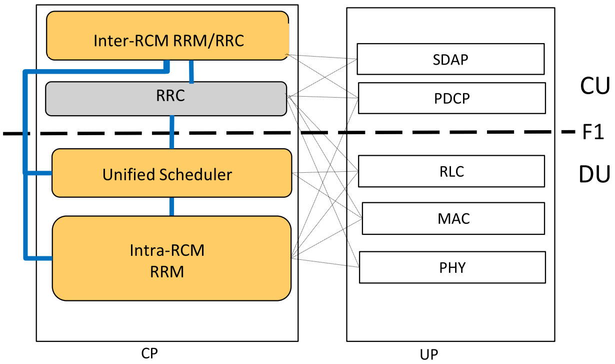


FIGURE 4. Exemplary functional deployment and interactions.

In particular, the CU-DU split option specified by 3GPP introduces a split between Packet Data Convergence Protocol (PDCP) and Radio Link Control (RLC) layer, as described in [15]. This means that Physical layer (PHY), Medium Access Control (MAC) and RLC will be located in the DU while PDCP and Service Data Adaptation Protocol (SDAP) plus Radio Resource Control (RRC) will be located in the CU, SDAP for the UP stack and RRC for the CP protocol stack. To support balancing of computational resources between CUs and DUs for gNBs, e.g., in a telco environment comprising centralized and edge clouds, the F1 interface as well as the Xn-C interface need to carry information about computational resource usage, such as for example CPU, memory, and network interface utilization. Figure 3 illustrates the 5G-RAN architecture and the CU-DU split of functionalities.

For allowing slice-awareness at RAN level, some principles have been discussed in [16], where an example illustration of the architecture enhancements is provided in Figure 4. The network slice-awareness in 5G RAN will strongly affect the RAN design and particularly the CP design, where multiple slices, with different optimisation targets, will require tailored access functions and functional placements to meet their target KPIs. On the other hand, the RAN deployment may provide some limitations on the efficiency of RRM due to the wireless channel, traffic load, and resource availability constraints, which may affect the overall performance (assuming numerous slices re-using the same RAN deployment). In particular, in dense urban HetNet scenarios, the signaling and complexity of RRM will be higher due to more signaling exchanges needed for passing RRM information to different entities. Moreover, the distribution of RRM functions in different radio nodes will provide new dependencies between RRM functions, which should be taken care of in order to optimize performance. In addition, in case of HetNet RAN deployments, non-ideal backhaul (with limited capacity and non-negligible latency) between access nodes (macro and small cells) will put some limitations on the RRM decisions and placement options to meet certain KPIs.

To ensure meeting the E2E slice requirements, assuming limited RCMs, which may be mapped to numerous slices, we introduce below a CP functionality framework, which is high required to allow for slice-tailored optimisation in RAN. In particular:

Intra-RCM RRM: Slice specific resource management and isolation among slices, utilizing the same RAN is an open topic which is currently investigated. In [17], the conventional management of dedicated resources can be seen as intra-slice RRM, which can be tailored and optimized based on slice specific KPIs.

Inter-RCM RRM/RRC: On top of Intra-RCM RRM, Inter-RCM RRM/RRC (which includes also Inter-slice RRM and slice-aware Topology RRM for wireless self-backhauling) is defined as the set of RRM policies that allow for flexible sharing, isolation, and prioritization of radio resources among slices or slice types in coarse time scales.. Inter-RCM RRM is an "umbrella" functional block to optimize the resource efficiency and utilization. In this direction, an Inter-RCM RRM mechanism is proposed in [18], where slice-aware RAN clustering, scheduler dimensioning, and adaptive placement of Intra-slice RRM functions is discussed in order to optimize performance in a dense heterogeneous RAN. Given the requirements of new access functions which can be tailored for different network slices, the distribution of RRM functionalities in different nodes is a key RAN design driver which allows for multi-objective optimisation in a multi-layer dense RAN. The adaptive allocation of such functions is also envisioned as key feature to cope with the dynamic changes in traffic load, slice requirements, and the availability of backhaul/access resources. To this end, one further Inter-slice/RCM RRM functionality is proposed in [19] which performs traffic forecasting of different slices and allocates resources to slices in a pro-active manner.

Topology RRM: This can be seen as another category of Inter-RCM RRM, mainly for distributed-RAN (D-RAN), where the resource allocation of wireless self-backhauling is essential to allow for joint backhaul/access optimisation [20]. Thus, Topology RRM can be tailored for different slices [21] in order to allocate backhaul resources among RCMs in a slice-tailored manner in order to avoid backhaul bottlenecks.

Unified Scheduler: This is an overarching MAC Scheduler, where different slice types share the same resources and dynamic resource allocation and slice multiplexing is required on top of RCM-specific MAC.

Based on this categorisation, an interesting aspect which may define the CP functionality requirements and the interface / signaling requirements between the CP functions is the functional split which is dependent on the CU - DU split options. In Figure 4, the possible placement of Inter-RCM and Intra-RCM RRM and RRC functionalities. Depending on the placement, the interface requirements might be different due to the time/resource granularity of the CP functionalities and their possible interconnections.

In literature, the abstraction of RAN control functionalities are under investigation from industry driven fora and alliances. In particular, O-RAN [3] is a new alliance which investigates the virtualization of access domain and considers the virtualization of control functionalities (RRC/RRM) to a newly defined RAN Intelligent Controller (RIC) which may reside either at the gNB or can be deployed for a cluster of gNBs. In this context, given the deployment and the functional requirements (real-time, non-real time, near-real-time) as well as the slice isolation policies, the above functional blocks can be either centralized at the CU (or above CU to a dedicated RAN controller) or distributed in DUs.

Data analytics could greatly benefit from such virtualized deployments, since the virtualization of control functions may impose signaling and complexity constraints especially for dynamic radio resource management mechanisms. The deployment of databases in central RAN nodes which run algorithms for predicting radio related context may support RRM functions so as to avoid relying on real-time measurements, while exploiting the benefits of centralized radio resource control and management.

B. 5G CN ARCHITECTURE

The key technological components of the 5G CN Architecture rely upon principles of *architecture modularisation, CP and UP separation, and* SBI. These are reflected in the SBA as mentioned earlier (specified since 3GPP Release 15) where the CP NFs are interconnected via a common SBI as shown in Figure 5. Compared to the traditional functional based network architecture design, SBA is expected to have the advantage of short role out time for new network features, extensibility, modularity, reusability, and openness. As outlined earlier, NWDAF is one key function within SBA, facilitating access to network data analytics. In Release 16, 3GPP SA2 started a new study Item, FS_eNA [22], to study enablers for Network Automation for 5G to further clarify the usage of data analytics capability in the network layer. FS_eNA envisions improving NWDAF scope via introducing use cases and solutions for supporting network automation deployment and the related framework (as shown in [11]) to collect/provide data analytics in relation to different NFs, AFs, and Management Functions (MFs), i.e., OAM.

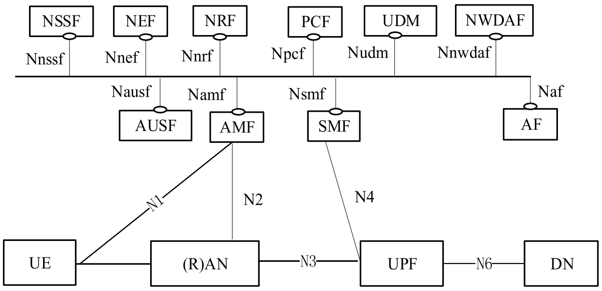


FIGURE 5. 3GPP SBA since Release 15 [11].

NWDAF reuses similar service exposure mechanisms as other 5G NFs (as described in 3GPP Release 15) for data collection and data analytics exposure from / to other NFs. There are also ongoing proposals to define a new service for unified data collection/analytics exposure from/to NFs/AFs.

The studied use cases in eNA study item include: NWDA-assisted QoS provision/traffic handling/customized mobility management/policy determination/QoS adjustment/5G edge computing/load (re-)balancing of NFs/determination of areas with oscillation of network conditions/slice SLA assurance/predictable network performance. NWDAF is also used for information retrieving from the application function, performance improvement and supervision of massive IoT terminals, prevention of various security attacks and UE driven analytics sharing.

In addition, a new use case on "UE-driven analytics sharing" has been proposed in this study item for enhancement of NWDAF and/or other NFs. In this use case, UEs are natural data collection points to gather more localized analytics within the network. Examples of data that a UE can provide are positioning information (e.g., collected from inertial or other sensors of a UE) or user profiling info (e.g., when a UE changes environment from outdoor to indoor or from vehicular to pedestrian mode). Such information may help the NWDAF to make more intelligent decisions on slice selection (e.g., to switch from a slice with more flexible resources to a resilient one or vice versa). As UEs can simultaneously connect to or switch across different slices (e.g. in case of mobility), they can have more prominent role for data preparation for the network to provide relevant localized contextual information and to identify earlier any changes in the network compared to the past intra-slice and/or inter-slice. Some key issues proposed and currently investigated for this use case include "How the NWDAF collects the UE's information" or "How the NWDAF uses the data provided by the UE to do analytics and provides the analytics information to other NFs".

One key consideration for allowing the inter-domain interaction of data analytics in 5GS is discussed in 3GPP SA5. For the interaction of NWDAF with OAM and RAN, the data collection from OAM may reuse the existing SA5 services. And how NWDAF provides the data analytics to OAM is still under discussion. Further details can be found in [23]. Figure 6 provides an overview of how analytics can be used across CN, RAN, and OAM to enable network automation.

C. 5G MANAGEMENT ARCHITECTURE

For Release 15, 3GPP SA5 has specified an architectural framework for telecom management that realizes a SBA approach. In this framework, a management service offers management capabilities that can be accessed by service consumers via a standardized service interface. Such management services include, for example, the performance management services, configuration management services, and fault supervision services. Consuming services may in turn produce (expose) these services to other consumers. Service producer and consumer may interact in a synchronous ("request-response") or asynchronous ("subscribe-notify") manner [24]. Within this framework, 3GPP SA5 introduced the MDAF that exposes one or multiple MDAS(s). Unlike an atomic function, an MDAS can exist at NF, network slice subnet, and network slice level. Deployment options for MDAS comprise centralized deployment (e.g., at a PLMN level) and domain-level deployments (e.g., RAN and CN network slice subnet instances, NSSIs). Domain MDAS provides domain-specific analytics, e.g., resource usage prediction in a CN or failure prediction in a subnet. A domain MDAF produces domain MDAS that is consumed by the centralized MDAF or another authorized MDAS consumers (e.g., infrastructure manager, network manager, slice manager, slice subnet manger, other 3rd party OSS). A centralized MDAS can provide E2E or cross-domain analytics service, e.g., resource usage or failure prediction in a network slice, optimal CN node placement for ensuring lowest latency in the connected RAN. A centralized MDAF produces centralized MDAS, and it is consumed by different authorized MDAS consumers.

3GPP SA5 also describes use cases for MDAS usage in [25], the technical specification on Performance Assurance. At this rather early stage, two major requirements have been specified: MDAS service producer has to allow its authorized consumer to request collection of (1) management analytical KPI(s) for network slice instances (NSIs) and NSSIs, and (2) management analytical KPI(s) for network(s). Details of the management analytical KPI(s) including, e.g., format, categorization and method/algorithm of calculations are yet to be defined. The uses cases for data analytic collection focus on the MDAS activity of creating the appropriate measurement jobs toward the monitored network resources. The MDAS producer first consumes the performance data reporting services to acquire the required measurements for NSI(s), NSSI(s), and NF(s). These form the basis to generate management analytical KPI(s) made available to other consumers, e.g., management services responsible for reporting the KPIs.

3GPP SA5 is also looking at the open source scenario for the data analytic topic. Specifically it started a study [26], on the Open Networking Automation Platform (ONAP) Data Collection, Analytics, and Events (DCAE), the module for

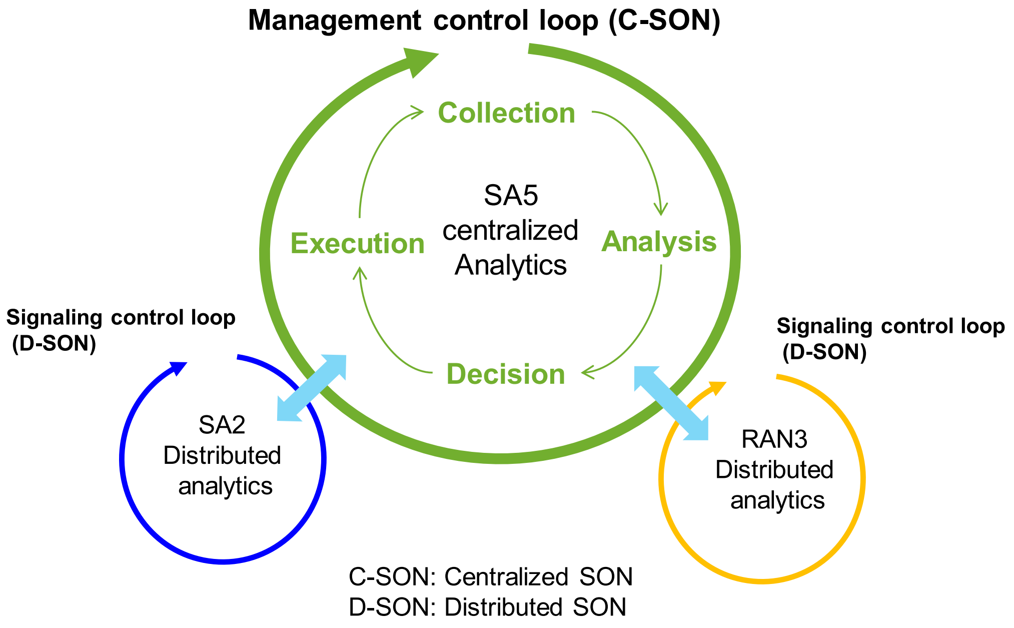


FIGURE 6. General framework for 5G network automation [23].

data collection and analytics. DCAE together with other ONAP components, gathers performance, usage, and configuration data from the managed environment. This data is then fed to various analytic applications, and if anomalies or significant events are detected, the results trigger appropriate actions. As a part of this study, 3GPP SA5 is comparing the data analytic approach and implementation of SA5 and ONAP to define new requirements for 3GPP SA5 MDAS or to give requirement to the ONAP consortium on DCAE.

V. INTEGRATED DATA ANALYTICS ARCHITECTURE

Taking into account the above discussions on the requirements for analytics in different parts of the 5GS and the current 5G architecture which supports the employment of data analytics in Core and Management domains, this section aims to propose an integrated analytics architecture to address the following challenges:

- Data Analytics in domains like RAN, CN, OAM domain, Application domain, data network (DN) is not yet fully specified in 5G architecture.
- The interactions between different analytics functions need to be discussed with particular examples and parameters, to better capture the necessary architecture enhancements in 5GS.
- The slice/service tailored E2E analytic functional parameterization and orchestration is not yet supported.

Therefore, we initially propose the functional blocks of an enhanced 5G architecture for performing multi-level analytics in different domains and for configuring / parameterizing analytics functionalities E2E in a slice-tailored manner. This involves the control, management, application, and data network domains, whereas the interconnection can be realized using a holistic SBA which covers all the aforementioned domains.

In this section, we discuss architecture enhancements and functional design considerations. Hence, the front-end is explicitly described as placeholder for employing analytics. The actual processing and data mining (e.g., what type of predictors and algorithms is used on top of these functionalities) and whether this involves multiple iterations and interaction between different entities is not shown, since this is an implementation specific aspect. Our intention is to prepare the grounds in 5G architecture for supporting analytics in multiple levels with different objectives, while these can be consumed by any authorized functionality is a slice-tailored manner.

In the context of cloud-ready telecommunications networks, 5G mobile NFs can be deployed as VNFs over geo-distributed cloud infrastructures with low effort. This decoupling of the computing and storage hardware from functionality embedded in software allows the easy deployment, maintenance, updating, and extension of 5GS functionality and offers unprecedented levels of network flexibility. Complementing network function virtualization (NFV), 3GPP has specified a SBA for the 5GS. It standardizes the 5G CN domain, mandating SBI to expose services of the CP functions, including NWDAF. Similarly, 3GPP has also followed a service-based approach in the Network Management domain, where Management Functions shall expose Management Services (e.g., Performance or Configuration Management service) to other consuming functions. Significant limitations currently comprise the lack of SBA support of NG RAN NFs as well as the service exposure and consumption across different network domains (e.g., RAN and CN) or across different planes,

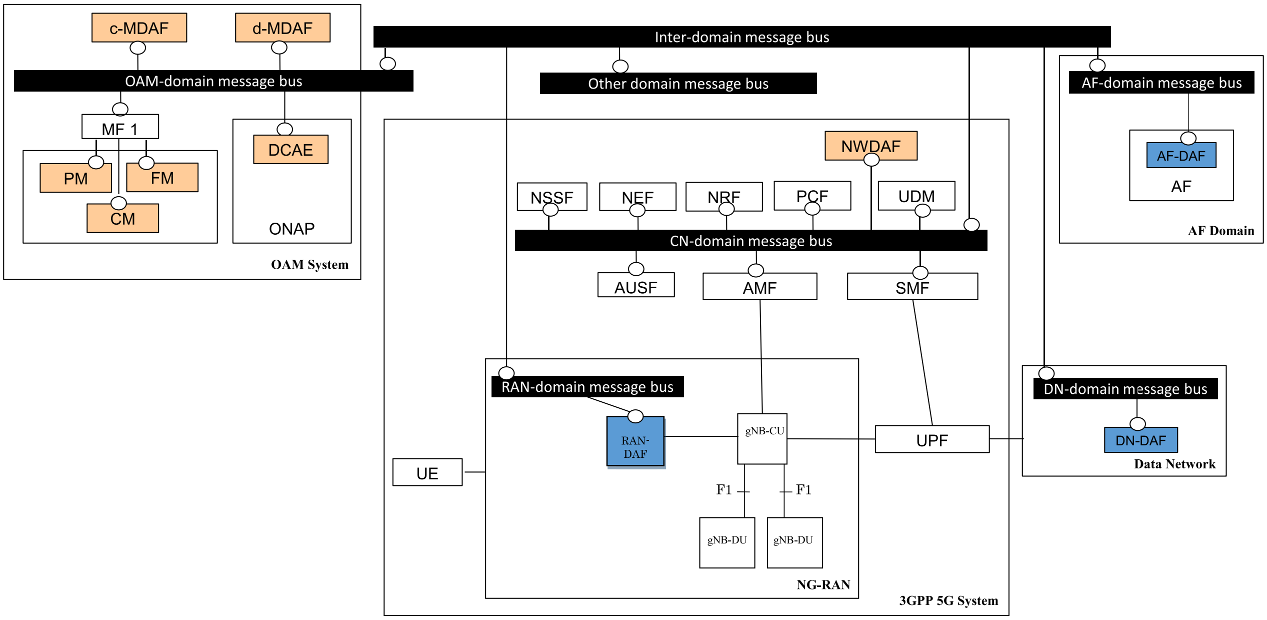


FIGURE 7. Integrated analytics architecture.

e.g., for interactions between CP and management plane services.

The necessity for new analytics functionality in 5GS may become reality, preferably using SBA, since both network operators and verticals may easily deploy analytics on demand. For example, analytics functions may be realized as (part of) a new AF, which can closely interact with, e.g., CN functions using SBI, or as CN/RAN functions which can interact with MDAF using the control-to-management interfaces, cf. Figure 7.

Apart from NWDAF and MDAF which were defined Section IV, the functionalities which can be defined as necessary parts of the E2E analytics design framework, as can be shown in Figure 8, are the following:

AF-DAF and DN-DAF: Outside the 3GPP 5GS, there are two relevant domains where additional data analytics functions can be deployed. In the DN, the network operator or the vertical can place functions that provide data related to service or performance of non-3GPP networks (e.g., metropolitan wide area networks, WANs) to other DAFs within 5GS or OAM domain. AFs or dedicated AF-DAFs can interact with CN-domain NWDAF, either via 3GPP Network Exposure Function (NEF) or via an inter-domain message bus as depicted in Figure 7. AF-DAFs enable the operator to deploy on demand new functionality customized for AF-domain requirements, or the vertical to perform analytics that can support the E2E service operation. This can prove highly beneficial for vertical industries like IIoT and V2X, where the vertical requires exposure of selected data from 3GPP network operation, a higher level of control of the network, as well as flexibility of deployment.

RAN-DAF: Real-time analytics are required for improving RAN NFs, like radio resource management. Since the RAN need to provide fast decisions, the analytics based on the processing of real-time measurements may need to stay local for optimizing performance dynamically. Also, there are the business aspects, which may involve different stakeholder among RAN, CN, and Management. So, the storage and analysis of radio-related measurement may be restricted to be abstracted to CN or OAM. An example deployment of such functionality is shown in Section VI, where more complex RAN deployments with CU-DU splits, better motivate for such functionality. Here, different options for performing RAN analytics may be examined. Either RAN-DAF will be a Control functionality at RAN or it can be a management/ SON functionality. With the proposed SBA as envisioned for both control and management functionalities, for both implementations of RAN-DAF, the interface will be via the inter-domain message bus interface.

Intra- and inter-domain message buses provide the functionality for registration, discovery, and consumption of services within a domain or across domains. Service registration and deregistration allows a service catalog function to maintain an updated list of services available for consumption. Service discovery functionality allows to retrieve available services, refer requesting consumers to them and provide the means to access them. Service consumption functionality allows consumers to invoke services, e.g., by automatically routing requests and responses between service consumer and producers. This may include platform-like functionality, such as, load balancing, failover, security, message delivery rules, or protocol conversion / adaptation, and exposure of services to the inter-domain message bus and its service catalog [27].

One of the major issues to be resolved in SBAs, such as the one depicted in Figure 7, comprises access control management, i.e., how are service consumers authenticated and authorized. For this purpose, [11] defines the Network Repository Function (NRF) that offers a service-based interface to allow NFs to do ''NF service registration'' by supplying the NF profile. The NF profile contains, among others, NF instance ID, type, network slice related Identifier(s), fully qualified domain name (FQDN) or IP address of NF, NF Specific Service authorization information, endpoint address(es) of instance(s) of each supported service, and names of supported services. The latter are used by NRF to provide service discovery to requesting NFs (incl. cross-PLMN service discovery), authorization information is used to grant (deny) authorization to the requesting NF to consume another service. This may be done through methods specified in [28]. In the case of the NWDAF, the so-called NF Service Authorization feature allows a consumer function to request a service, e.g., receiving the requested data from the NWDAF service. The details of the authorization procedure based on NF profile are covered in [11] and [28]. Moreover, due to roaming agreements and operator policies, a NF Service Consumer shall be authorized based on UE/subscriber/roaming information and NF type.

While these procedures provide the fundamental means to securely provide services within and across domains, 3GPP largely restricts service consumption to CN CP NFs. The solution proposed here reuses and extends 3GPP mechanisms to allow inter-domain exposure and consumption, taking the use case of inter-domain data analytics services (DASs). Therefore, this article proposes to enhance current ''per service'' authorization for DASs towards per ''data element'' (DE) authorization, using three core features:

(1) Different rights of access to data element ('permission'): A DE shall have multiple access levels which entities may be entitled to.
(2) Differentiation among consumers that can access a service (role-based model): Each consumer shall be grouped to belong to at least one class of consumers (''role''), based on the consumer profile provided. This provides the identity feature or role of the consumer to be used for authorization granting.
(3) Definition of DE access rights by the service provisioning function (e.g., the NF that is the actual producer or ''owner'' of the DE).

The service provisioning function of a DE identifies the role(s) that are entitled to consume its services and the level at which each role can access the data. The source NF shall define the access rights either in registering a recurring data stream or for each single data element that it publishes (e.g., to NWDAF or a repository), respectively.

Given the aforementioned types of analytics and the proposed architecture enhancements, Table 1 provides some exemplary functionalities which can be defined and configured based on different slice requirements and network conditions.

TABLE 1. Analytics functionality placement and classification.

	Parameter	Type	Placement	Time-scale
A. UE-related parameters	Mobility	Prescriptive Analytics	RAN-DAF	Real-time
		Descriptive Analytics , Predictive Analytics	AF-DAF	
	Interference level	Predictive Analytics, Prescriptive Analytics	RAN-DAF	Real-time
	UE QoS	Diagnostic Analytics, Predictive Analytics	NWDAF	Real-time / non-real time
		Prescriptive Analytics	AF-DAF	
B. Network-related parameters	Radio Resource Situation (conditions, usage, availability)	Diagnostic Analytics, Predictive Analytics, Prescriptive Analytics	RAN-DAF	Real-time / non-real time
	Traffic / Load Situation	Diagnostic Analytics, Predictive Analytics	NWDAF	Non-real time
		Prescriptive Analytics	MDAF	
	Backhaul Conditions / Availability (e.g. BS neighborhood change)	Descriptive Analytics , Predictive Analytics	RAN-DAF	Non-real time
		Diagnostic Analytics,	NWDAF	
		Prescriptive Analytics	MDAF	
	Cell Density	Diagnostic Analytics, Predictive Analytics	NWDAF	Non-real time
		Prescriptive Analytics	MDAF	
	Network QoS	Diagnostic Analytics, Predictive Analytics	RAN-DAF, NWDAF	Real-time / non-real time
		Prescriptive Analytics	AF-DAF	
C. Service-related Parameters	New / Modified Slice	Prescriptive Analytics	UE, AF-DAF	Real-time / non-real time
	NW assistance (for V2X)	Predictive Analytics	RAN-DAF, NWDAF	Real-time / non-real time
		Prescriptive Analytics	AF-DAF	
	UE Route / Trajectory (for V2X)	Prescriptive Analytics	AF-DAF	Real-time / non-real time
	Level of Automation Change (for V2X)	Prescriptive Analytics	UE, AF-DAF	Real-time / non-real time
D. Management-related parameters	PM	Descriptive Analytics, Diagnostic Analytics	MDAF	Non-real time
	FM	Descriptive Analytics, Diagnostic Analytics	MDAF	Non-real time

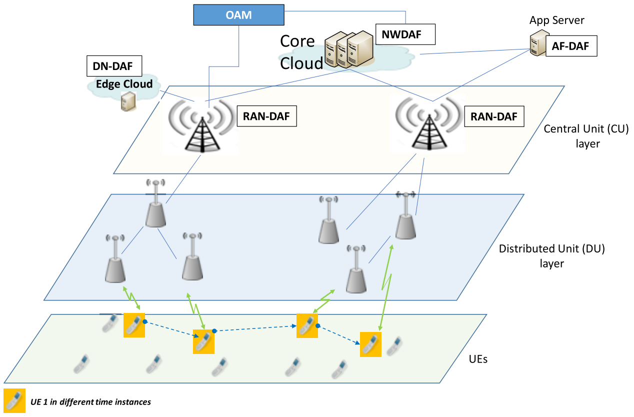


FIGURE 8. Illustration of exemplary case study.

VI. CASE STUDY FOR E2E ANALYTICS DESIGN FOR RAN OPTIMIZATION

This section presents a case study which focuses on the proposed RAN-DAF and AF-DAF functional elements. It particularly shows the interactions between them, and how RAN-DAF can enhance an RRM functionality which may reside in CU or DU. An exemplary deployment is shown in Figure 8.

Initially, OAM provides RAN configuration (functional and RRM policies) to RAN. These policies may be pre-configured or can be triggered by MDAF or by NWDAF with respect to possible changes required at RAN for one or multiple slices. The RRM policies as specified in 3GPP SA5 can be seen as abstract guidelines from the management system to RAN. RAN is the domain which will enforce RRM based on real-time information. In regards of slicing, these policies may include the split of resources between slices or the level of isolation in RAN.

The 3$^{\text{rd}}$ party analytics, either deployed as AF or as Local DN (for multi-access edge computing (MEC)-enabled RAN nodes), may provide the prediction of the UE routes. This particular input to RAN-DAF may be considered as a realistic scenario for V2X use cases, since the network may not be aware of the precise location of the UEs. In this scenario, the UE routes/trajectories can be predicted at the V2X application server or client; hence, the analytics at the 3$^{\text{rd}}$ party domain (cloud or application function) allows the 5G network efficient resource management based on predicted load due to the expected UE locations while having ongoing sessions.

For the predicted traffic load and UE mobility, Liotou *et al.* [30] propose a solution which models the user mobility using Self-similar Least-Action Walk (SLAW) model [31], derived by empirical studies of real-life human-walk traces. SLAW exploits the property of "gravity points" or so-called "clusters", i.e., popular areas where users tend to move and accumulate with high probability, providing a realistic outlook of network traffic. In particular, let *G(V, E)* a graph consisting of the access nodes, starting from the CU towards the UEs traversing one or more DUs via single or multiple hops (as can be seen in Figure 8). The predicted traffic demand of a user $u \in U$ over each link $e \in E$ *is defined* as $Demand_t(e, u)$, to identify the rate requirement that will contribute to the load of that link for each time instance, if this link is used to carry traffic. Each link e has an upperbound capacity $c_e, \forall e \in E$, that corresponds to the maximum rate over that link. Furthermore, the expected traffic load which corresponds to the summation of all the user traffic traversing it for each time frame is defined as follows:

$$\sum\nolimits_{u\in U} a_{u(e=\{i,m\},t)} Demand_t\left(e,u\right), \quad \forall t,\ \forall i \neq m,\ \forall u \in U.$$

Here, $a_{u(e=\{i,m\},t)}$ is a binary variable which can be derived by the output of the mobility model as in [30], which determines the actual time frame(s) during which each user receives/transmits from/to a DU $m \in M$ ($M \subseteq V$ depicts the set of all DUs), based on the assumption that the UE may reside at this DU for a given time. This variable is 1 if the user u resides at DU m at time instance t and only if m is one node at edge e; otherwise it is 0. Using the notion of link capacity and link load, a new parameter, which captures the utilization of the link is:

$$c_{e,t} = \frac{\sum\nolimits_{u\in U} a_{u(e=\{i,m\},t)} Demand_t\left(e,u\right)}{c_e}, \quad \forall t \in T, \forall e \in E \tag{1}$$

where $c_{e,t}$, is the normalized cost function of each link, which captures how much percentage of utilization of the link is expected (from the CU to the destination DU serving the UEs). The expected load at the access nodes, together with the information on the UE locations at each time instance, may thus be used by RAN-DAF to estimate the access congestion probability (and the expected qualities of the allocated links) per time window, considering the potential interference from surrounding active links.

A. ANALYTICS AT RAN-DAF USING REAL MEASUREMENTS

The mobility and traffic prediction can be outputted to RAN-DAF as the aforementioned utilization metric $c_{e,t}$ for the link utilization by traffic corresponding to different UEs, as well as the user expected locations. This in-turn can help estimating the signal quality of the UE(s) over the expected path. The derivation of the expected signal-to-interference-plus-noise ratio (SINR) is key for allowing pro-active adaptations to ensure meeting the KPI requirements. In particular, while knowing the traffic expectation in a given area, we need to also know the expected channel conditions, which may limit the performance gains by access factors like interference, overload/congestions, and coverage gaps.

The expected SINR can be formulated as:

$$\widetilde{SINR_{u,m,t}} = \frac{P_{u,m,t}G_{u,m,t}\tilde{L}\left(d^{\rho}_{u,m}\right)}{\bar{I}+\eta}, \forall u \in U, \forall m \in M, \forall t \in T, \tag{2}$$

where $P_{u,m,t}$ and $G_{u,m,t}$ are the transmit power and channel gain per timeslot from m to u respectively. $\tilde{L}\left(d^{\rho}_{u,m}\right)$ is the distance dependent path-loss based on the expected user location (i.e., d is the distance between u, m and ρ is the path loss exponent). Furthermore, $\bar{I}$ is the average interference (due to power leakage by other sources) and η is the thermal noise.

Based on the SINR expectation, the expected link capacities can be calculated; and can be generally expressed for each time instance as:

$$\begin{aligned}&\widetilde{R_{u,m}}\left(\overline{c_{e=\{i,m\}}}\right)\\ &\quad = E\left[\overline{c_{e=\{i,m\}}}\sum\nolimits_{t=1}^{T} a_{u(e=\{i,m\},t)}BW_u log_2\left(1+\widetilde{SINR_{u,m,t}}\right)\right],\\ &\qquad\qquad \forall i \neq m \in M,\end{aligned} \tag{3}$$

where $\overline{c_{e=\{i,m\}}}$ accounts for the mean utilization factor, averaged over T timeslots, and BW_u is the bandwidth allocated per user (can be variable given the scheduling policies corresponding to the number of UEs served by a DU). Hence, the expected capacity of the access link, takes into account the expected SINR based on the estimated user location, as well as the expected average utilization factor of the serving DU, which may limit the maximum allocated resources per cell (based on the expected aggregated demand).

Figure 9 shows an exemplary measurement campaign in Leipzig, Germany. We evaluate the downlink SINR in a vehicular UE associated to a local cellular LTE network. It is

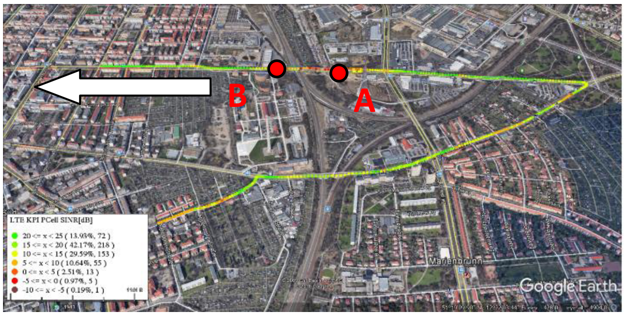

B
A
LTE KPI PCell SINR[dB]
20 <= x < 25 (13.93%, 72)
15 <= x < 20 (42.17%, 218)
10 <= x < 15 (29.59%, 153)
5 <= x < 10 (10.64%, 55)
0 <= x < 5 (2.51%, 13)
-5 <= x < 0 (0.97%, 5)
-10 <= x < -5 (0.19%, 1)
Google Earth

FIGURE 9. Measurement route in Leipzig.

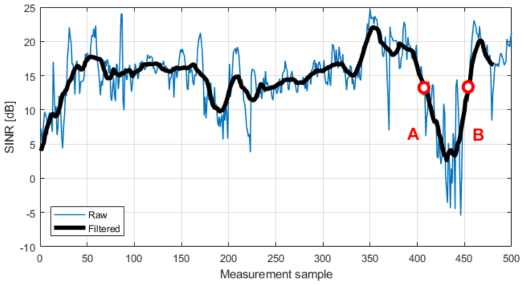


FIGURE 10. SINR measurement results.

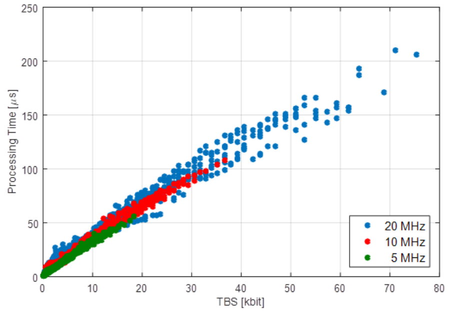


FIGURE 11. Channel encoding processing time evaluation.

assumed that the raw SINR measurements are periodically reported to the RAN-DAF where they are processed.

This scenario shows a critical region with low SINR levels between waypoint A and waypoint B. Such an area could for example become critical in terms of reliability in case of vehicular communications for autonomous driving. An excerpt of the corresponding SINR measurements samples is shown in Figure 11. After filtering the raw SINR samples in the RAN-DAF, for example by means of a moving average filter, they could be used for building prediction models for critical street segments within the AF-DAF. Such a prediction model could then be used for mobility route optimizations for vehicular UEs.

An additional use case could be the resource requirement prediction for sites within the mobile radio network, comprising both radio resources at the air interface and computational resource for specific VNFs within the RAN depending on UE positions.

Regarding the latter, Figure 10 shows an exemplary channel encoder processing time evaluation conducted with a virtualized eNB based on srsLTE [32]. The processing time which represents a metric for the computational resource utilization in a virtualized environment depends on the system bandwidth and the scheduled transport block size (TBS) within a transmission time interval. Since the selected TBS furthermore depends on the SINR due to adaptive modulation and coding in combination with specific multiple-input multiple output (MIMO) transmission schemes, the processing time for specific UE positions can be predicted by combining the SINR prediction model from Figure 10 with processing time models derived from the measurements shown in Figure 11. Such a model can be provided by the AF-DAF to the RAN-DAF in order to support computational resource aware scheduling decisions.

B. USE OF DATA ANALYTICS FOR RAN OPTIMIZATION FOR V2X SERVICES

The expectation of SINR and the calculation of the expected KPI (e.g. link expected capacity) as provided by the database which may reside in a CU, taking into account the expected UE routes, may be beneficial for allowing RAN to optimize the resource allocation in a multi-cell environment.

Here, two possible pro-active actions can be performed by RAN to ensure meeting the requirements, targeting vehicular scenarios which include safety related applications and due to high mobility the real-time resource optimization may be challenging:

- Dynamic adjustment of 5G QoS indicator (5QI) attributes for UEs with good or bad channel conditions, to avoid service discontinuity due to poor coverage or cell-edge channel quality.
- Pro-active adjustment of Resource Block (RB) selection using multi-cell joint scheduling to avoid interference, while not requiring real-time measurements from UEs at CU.

One of the ongoing discussions in 3GPP SA2 eV2X study is enabling RAN to get from CN multiple QoS levels which can be mapped to a session. This will allow RAN to adapt fast the QoS level if possible QoS downgrade is monitored. This would be beneficial for safety-critical and low latency V2X services, since the CN does not need to perform actions to adapt the QoS level.

An example is the adaptation of Packet Delay Budget (PDB) attribute of 5QI, which could be reduced for UEs with good channel quality and increased for cell-edge UEs. This may be well applicable to adapting the PDB for uplink and downlink traffic which may have the same QoS class, taking into account the channel difference between an UL and downlink (DL) user.

As can be shown in Figure 12, assuming an example of Vehicle 1 (with low channel quality indicator, 5QI) and Vehicle 2 (with high 5QI) and Delay Budget of 5ms for UL/DL. UL needs more transmission time intervals (TTIs)

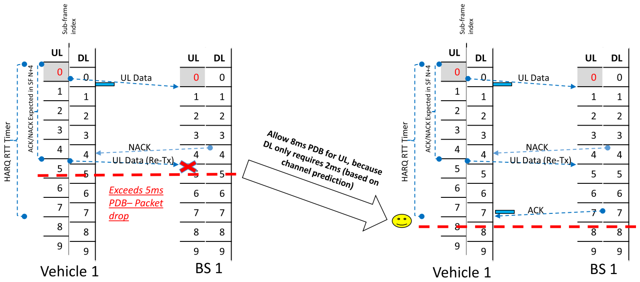

Sub-frame index
UL
DL
HARQ RTT Timer
ACK/NACK Expected in SF N+4
UL Data
NACK
UL Data (Re-Tx)
Exceeds 5ms PDB– Packet drop
Vehicle 1
BS 1
Allow 8ms PDB for UL, because DL only requires 2ms (based on channel prediction)
Sub-frame index
UL
DL
HARQ RTT Timer
ACK/NACK Expected in SF N+4
UL Data
NACK
UL Data (Re-Tx)
ACK
Vehicle 1
BS 1

FIGURE 12. Impact of dynamic 5QI adjustment on reliability.

Algorithm 1 V2V-Tailored UL/DL Pattern Selection

```
Step 0: Create 7 different vectors for TDD patterns
        (20/80 TD1 = [], 30/70 TD2 = []; 40/60 TD3 = [];
        50/50 TD4 = []; 60/40 TD5 = []; 70/30 TD6 = [];
        80/20 TD7 = []
Step 1: $\forall$ V2V path with transmitting UE x of BS i and
        receiving UE y of BS j: Use RAN-DAF to get the
        expected Channel Gain or SINR (based on RSRP)
        of link 1, (Channel(x,i)) and of link 2 (Channel (j,y)
Step 2: Compute and store the Ratio of expected SINR
        between link 1 and 2 and get the closest integer for
        both cases.
Step 3: Based on the Ratios, map the path (s) to the relevant
        TDD pattern and store the link mapping to the TTIs
Step 4: Go to Step 1 till no path can be added
```

Algorithm 2 Graph-Based Resource Allocation

```
- Set FL as [#Links x #TTI] matrix from Algorithm 1
- Set a color set Color and maximum number of colors Cmax
and Clist={}
for TTI=1:T
    if FL(1:links, TTI)==y $\leq$ Cmax
        Set randomly $y \in$ Colors different colors for the
        links connecting to TTI
    end if
        Store color indices for all links for TTI in a matrix as:
        Coloring(Link, Color Index, TTI)
end for
for color_index=1:Cmax
    CList(color_index)=Coloring (1:links,color_index,1:T)
end for
for bands=1:total RBG and color_index=1:Cmax
    Map bands to CList(color_index) that maximizes
    sum-rate
end for
end for
```

for re-transmissions than DL, otherwise the packets will be dropped. By this adaptation, we manage to enhance reliability by assigning more TTIs for UL so as to allow successful re-transmission to avoid dropping packets.

The use of analytics for the channel gain difference for UL and DL traffic, based on predicted routes and traffic demand, will allow for fast PDB adjustment in order to meet the KPIs.

On top of the of the PDB adjustment, for resource optimization, we can perform multi-cell scheduling to adapt the RB allocation. This will require a lot of signaling for getting real-time measurements from UEs and coordination between gNBs and signaling over F1 (CU-DU interface). If we assume also, UL and DL scheduling in dynamic-TDD (D-TDD) scenarios, the signaling and complexity can be very high.

The awareness of the expected SINR distribution over a time window, may exploit the benefits of sophisticated scheduling while keeping the signaling effort low. This may allow for virtualization of RRM at the cloud, since the control signaling latency constraints can be relaxed.

Below a two-step algorithm is proposed for the UL/DL TTI adaptation and joint scheduling in a multi-cell environment taking into account the V2X scenario (including UL and DL parts of a vehicle-to-vehicle (V2V) indirect session):

- A1. V2V-tailored TDD configuration: The objective is to find the optimal UL/DL Configuration per V2V session based on expected Channel Gain/SINR Difference between users of the path.
- A2. Joint UL/DL Resource Allocation per TTI to minimize interference. This uses a graph-coloring based solution, which aims to minimize interference by finding conflicts at each TTI so as to avoid using the same Resource Block Group (RBG). More information and evaluation of Algorithm 2 for a D-TDD scenario can be found in [33].

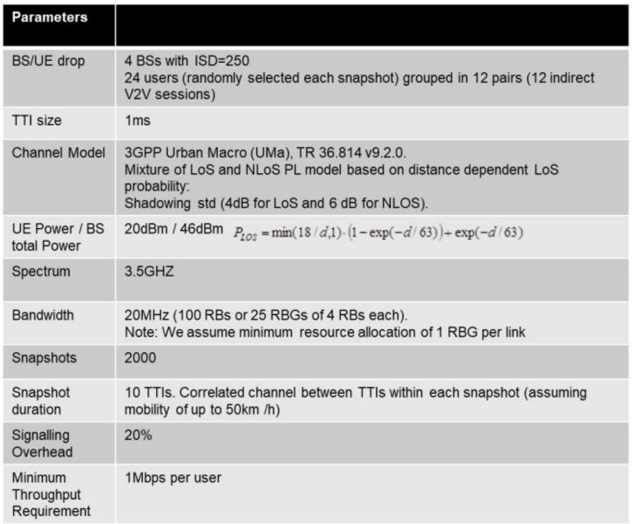
The Algorithms 1 and 2 are shown in more detail below:

We performed Monte-Carlo Simulations for evaluating the above considerations (adapting PDB of the 5QI and on top of that adapting the resource allocation) for a scenario consisting of low-latency V2V services which are connected via Uu. The simulation parameters are captured in Table 2.

The evaluation shows the trend of E2E V2V throughput (which is the minimum of UL and DL counterparts of the V2V indirect path). The benchmark is the uncoordinated scheduling per BS (using PF intra-cell scheduling) and Fixed UL and DL patterns (1:5 TTI UL and 6:10 TTI DL). The first proposal is the fixed RB allocation, while the TTI is adapted based on the expected channel difference (Algorithm 1); however, there is no coordination between BSs for the RB allocation to minimize interference. Additionally, proposal 2 provides the adaptive pro-active allocation of resources, while the TTIs for UL and DL are also adapted (as proposed in Algorithms 1 and 2).

As can be seen in Figure 13, our Proposed Algorithms 1 and 2 outperform the benchmark in V2V E2E throughput. When Algorithm 2 is employed together with Algorithm 1, the gain is higher than Algorithm 1 and uncoordinated PF scheduling (16% at the median of the CDF) and no gain is observed at 10th percentile (which means that the gain is more for the on average “good quality” V2V paths). So, if both Algorithms 1 and 2 are employed, the gain is significant; however without the use of analytics, the com-

TABLE 2. Evaluation parameters.



Parameters	
BS/UE drop	4 BSs with ISD=250 24 users (randomly selected each snapshot) grouped in 12 pairs (12 indirect V2V sessions)
TTI size	1ms
Channel Model	3GPP Urban Macro (UMa), TR 36.814 v9.2.0. Mixture of LoS and NLoS PL model based on distance dependent LoS probability: Shadowing std (4dB for LoS and 6 dB for NLOS).
UE Power / BS total Power	20dBm / 46dBm $P_{LOS} = \min(18/d,1)\cdot(1-\exp(-d/63))+\exp(-d/63)$
Spectrum	3.5GHZ
Bandwidth	20MHz (100 RBs or 25 RBGs of 4 RBs each). Note: We assume minimum resource allocation of 1 RBG per link
Snapshots	2000
Snapshot duration	10 TTIs. Correlated channel between TTIs within each snapshot (assuming mobility of up to 50km /h)
Signalling Overhead	20%
Minimum Throughput Requirement	1Mbps per user

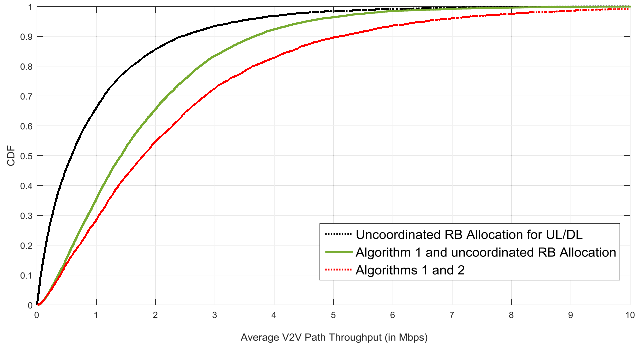


FIGURE 13. Cumulative distribution function (CDF) of V2V Session Throughput.

plexity and signaling costs to perform this in a centralized entity would be quite challenging (especially when assuming CU-DU splits are non-ideal backhaul). On the other hand, only the QoS adaptation, and in particular the dynamic setting of PDB for UL and DL may require less signaling effort since it happens less dynamically and the granularity and confidence interval of analytics could be relaxed.

VII. OPEN RESEARCH DIRECTIONS

The proposed architecture can be seen as a placeholder for integrated analytics and interactions among different domains, covering also different stakeholders involved for different services. Some open directions, which are key for the realization of such architecture, are the following:

- The actual analytics functionality which involves the data mining and processing may be implemented using different algorithms. In this article, the focus was on architecture and RAN-centric analytics; however, a potential future work is the comparison of predictors for different circumstances taking into account multiple levels of interacting analytics (e.g., using online and offline analytics and possible multiple iterations among different nodes).
- It is assumed that full SBA is used for multi-level analytics, however this may not be possible if we assume parts of the network to rely on different technologies (e.g., LTE and WiFi). In this context, further investigation is needed on how backward compatibility is supported.
- Exposure capabilities is a factor which can provide limitations and opportunities. Especially for AF and DN driven analytics, it should be noted that the exposure via NEF needs to be supported, and also this may affect the signaling delays when analytics of different levels require frequent signaling for performing prediction jointly.
- RAN domain is a special case, which currently is under discussion in 3GPP whether new analytics functionality is required or not. One particular aspect is how the RAN-DAF would be defined in 3GPP, e.g., as control or management functionality. This will strongly affect the interfacing to other domains like OAM (e.g., ONAP/network management/virtualization MANO) as well CN and application domain.
- Complexity of multi-level analytics is one factor which needs to be further discussed, especially if this affects real-time decisions (e.g., RRM). The decision point as well as the requirement for offline and online analytics, as well as the iterations required for optimization should be discussed further. For very dynamic decisions, it is usually argued whether analytics on channel qualities may be useful and accurate in millisecond time scale. This particular point needs to further investigated.
- One further open direction is the role of analytics and the impact on the 5GS. Is it going to be an essential feature or supportive/back-up feature for improving the network efficiency? For example, if prediction on the resource availability or network QoS fails, this may have impact on the whole system. Hence, we need to ensure 100% function availability/reliability when decisions are based on analytics.
- Furthermore, one additional direction is the security/integrity aspect of analytics. How can the operator ensure trustworthiness and data integrity of analytics from the terminal, application or domain? Flawed or malicious data analytics or non-accurate data can cause significant issues to the network, if it relies on them for taking actions.
- Finally, one key direction is the confidence intervals of analytic results and how this can be calculated if analytics are performed in a distributed manner. The exploitation of the benefits of prediction strongly depends on how accurate analytics can be, and this strongly depends also on the predictors/algorithms used for mining the data and process them.

VIII. CONCLUSIONS

The article has proposed an enhanced 5G architecture which allows for end-to-end support of data analytics. In particular, the requirements for data analytics in order to improve the operation in different domains have been explicitly presented. Subsequently, the enhancements of the current architecture of the 3GPP 5GS have been highlighted and key design directions have been noted. Finally, an enhanced SBA that seamlessly integrates data analytics functionality is proposed. This architecture incorporates several new functional blocks in application, DN, and RAN domains so as to allow both 3rd party and local data analytics, e.g., to perform fully customized and machine learning-supported service operations. To showcase the benefits of the new architecture, we provided a case study which uses both application and RAN analytics in order to optimize resource allocation in cases when the mobility of the users as well as the signal degradation in certain cell areas can be predicted. Finally, we have highlighted open research directions toward the realization of the proposed enhanced 5G architecture.

ACKNOWLEDGEMENT

This work has been performed in the framework of the H2020 project 5G-MoNArch funded by the EU. The views expressed are those of the authors and do not necessarily represent the project.

REFERENCES

[1] K. Samdanis, X. Costa-Perez, and V. Sciancalepore, "From network sharing to multi-tenancy: The 5G network slice broker," *IEEE Commun. Mag.*, vol. 54, no. 7, Jul. 2016.

[2] I. Afolabi, T. Taleb, K. Samdanis, A. Ksentini, and H. Flinck, "Network slicing and softwarization: A survey on principles, enabling technologies, and solutions," *IEEE Commun. Surveys Tuts.*, vol. 20, no. 3, pp. 2429–2453, 3rd Quart., 2018.

[3] *O-RAN Alliance*. Accessed: 2018. [Online]. Available: https://www.o-ran.org/resources/

[4] J. Hagerty, R. L. Sallam, and J. Richardson. (2012). Magic Quadrant for Business Intelligence Platforms. Gartner. [Online]. Available: http://www.microstrategy.com/download/files/whitepapers/open/gartnermagic-quadrant-for-bi-platforms-2012.pdf

[5] *Service Requirements for Next Generation New Services and Markets, v16.5.0*, document TS 22.261, 3GPP, Sep. 2018.

[6] *Study on Communication for Automation in Vertical Domains (CAV), v16.1.0*, document TR 22.804, 3GPP, Sep. 2018.

[7] *Study on Enhancement of 3GPP Support for 5G V2X Services, v16.1.1*, document TR 22.886, 3GPP, Sep. 2018.

[8] *Study on Application Layer Support for V2X Services, v16.0.0*, document TR 23.795, 3GPP, Sep. 2018.

[9] *NR; Overall Description, Stage-2, v15.3.1*, document TS 38.300, 3GPP, Oct. 2018.

[10] P. Marsch, O. Bulakci, O. Queseth, and M. Boldi, *5G System Design: Architectural and Functional Considerations and Long Term Research*. Hoboken, NJ, USA: Wiley, 2018.

[11] *System Architecture for the 5G System, v15.3.0*, document TS 23.501, 3GPP, Sep. 2018.

[12] *New Study on Self-Organizing Networks (SON) for 5G, v.0.2.0*, document TR 28.861, 3GPP, Dec. 2018.

[13] *Study on RAN-Centric Data Collection and Utilization for LTE and NR*, document RP-182105, 3GPP, Sep. 2018.

[14] E. Pateromichelakis, J. Gebert, T. Mach, J. Belschner, W. Guo, and N. P. Kuruvatti, "Service-tailored user-plane design framework and architecture considerations in 5G radio access networks," *IEEE Access*, vol. 5, pp. 17089–17105, Aug.

[15] *NG-RAN; Architecture Description, v 15.1.0*, document TS 38.401, 3GPP, Apr. 2018.

[16] S. Redana and O. Bulakci, "View on 5G architecture v2.0," 5G-PPP WG Archit., EU, 5G Architecture White Paper, Dec. 2017. [Online]. Available: https://5g-ppp.eu/wp-content/uploads/2018/01/5G-PPP-5G-Architecture-White-Paper-Jan-2018-v2.0.pdf

[17] F. Z. Yousaf and T. Taleb, "Fine-grained resource-aware virtual network function management for 5G carrier cloud," *IEEE Netw*, vol. 30, no. 2, pp. 110–115, Mar./Apr. 2016.

[18] E. Pateromichelakis and C. Peng, "Selection and dimensioning of slice-based RAN controller for adaptive radio resource management," in *Proc. IEEE Wireless Commun. Netw. Conf. (WCNC)*, Mar. 2017, pp. 1–6.

[19] V. Sciancalepore, K. Samdanis, X. Costa-Perez, D. Bega, M. Gramaglia, and A. Banchs, "Mobile traffic forecasting for maximizing 5G network slicing resource utilization," in *Proc. IEEE Conf. Comput. Commun. (INFOCOM)*, May 201, pp. 1–9.

[20] Y. Li, E. Pateromichelakis, N. Vucic, J. Luo, W. Xu, and G. Caire,, "Radio resource management considerations for 5G millimeter wave backhaul and access networks," *IEEE Commun. Mag.*, vol. 55, no. 6, pp. 86–92, Jul. 2017.

[21] E. Pateromichelakis, K. Samdanis, Q. Wei, and P. Spapis, "Slice-tailored joint path selection & scheduling in mm-wave small cell dense networks," in *Proc. IEEE Global Commun. Conf.*, Dec. 2017, pp. 1–6.

[22] *Study of Enablers for Network Automation for 5G, v.2.0.0*, document TR 23.791, 3GPP, Dec. 2018.

[23] *Discussion About How SA5/ SA2 / RAN3 Could Work Together to Guarantee Network Slice SLA*, document S5-186486, Oct. 2018.

[24] *Management and Orchestration; Architecture Framework, v15.0.0*, document TS 28.533, 3GPP, Sep. 2018.

[25] *Management and Orchestration; Performance Assurance, v2.2.0*, document TS 28.550, 3GPP, Dec. 2018.

[26] *Study on Integration of Open Network Automation Platform (ONAP) at a Collection, Analytics and Events (DCAE) and 3GPP Reference Management Architecture, v1.0.0*, document TR 28.900, 3GPP, Dec. 2018.

[27] *Zero-Touch Network and Service Management (ZSM), Reference Architecture, V0.7.1*, document GS ZSM 002, ETSI, Nov. 2018.

[28] *5G System; Technical Realization of Service Based Architecture, Stage 3, v15.1.0*, document TS 29.500, 3GPP, Sep. 2018.

[29] *Procedures for the 5G System, v15.3.0*, document TS 23.502, 3GPP, Sep. 2018.

[30] E. Liotou, K. Samdanis, E. Pateromichelakis, N. Passas, and L. Merakos, "QoE-SDN APP: A rate-guided QoE-aware SDN-APP for HTTP adaptive video streaming," *IEEE J. Sel. Areas Commun.*, vol. 36, no. 3, pp. 598–615, Mar. 2018.

[31] K. Lee, S. Hong, S. J. Kim, I. Rhee, and S. Chong, "SLAW: Self-similar least-action human walk," *IEEE/ACM Trans. Netw.*, vol. 20, no. 2, pp. 515–529, Apr. 2012.

[32] *Software Radio Systems*. Accessed: 2018. [Online]. Available: http://www.softwareradiosystems.com/tag/srslte/

[33] E. Pateromichelakis and K. Samdanis, "A graph coloring based inter-slice resource management for 5G dynamic TDD RANs," in *Proc. IEEE Int. Conf. Commun.(ICC)*, Kansas City, MO, USA, May 2018, pp. 1–6.

EMMANOUIL PATEROMICHELAKIS received the M.Sc. and Ph.D. degrees in mobile communications from the University of Surrey, U.K., in 2009 and 2013, respectively, where he was a Research Fellow with ICS, from 2013 to 2015. Since 2015, he has been a Senior Researcher with the Europe Standardization and Industry Development Department, Huawei GRC, focusing on 5G and beyond solutions with active participation in 5G EU Projects (METIS II, 5G-Xhaul, mm-MAGIC, and 5G-MoNArch) and 3GPP standardization activities. Part of this work has been published in several conferences and journal papers, and he has also filed more than 20 patents on 5G related topics.

FABRIZIO MOGGIO received the master's degree in electrical engineering, with specialization in telecommunications, from the Politecnico of Turin, Italy, in 1997. Since 1997, he has been with the Innovation Department, Telecom Italia. Since 2017, he has also been with Core Network innovation with a specific focus on 5G. He has also been the 3GPP delegate of Telecom Italia for the Telecom Management Group (SA5), since 2017.

CHRISTIAN MANNWEILER received the M.Sc. (Dipl.-Wirtsch.-Ing.) and Ph.D. (Dr.-Ing.) degrees from the Technische Universität Kaiserslautern, Germany, in 2008 and 2014, respectively. Since 2015, he has been a member of the Network Automation research group, Nokia Bell Labs, where he has been working in the area of cognitive network management and SON for 5G systems. He has worked in several nationally and EU-funded projects covering the development of cellular and industrial communication systems. He has authored or co-authored numerous articles and papers on wireless communication technologies and architectures for future mobile networks.

PAUL ARNOLD received the Dipl.-Ing. degree in information and communication techniques from the University of Applied Sciences Mittelhessen, Friedberg, Germany. He has been with Deutsche Telekom, since 2007, where he has been involved in several Horizon 2020 EU funded projects such as E3, METISII, 5G-NORMA, and 5G-MoNArch. He is currently a Research Engineer with Deutsche Telekom AG, Darmstadt, Germany. His research interests include wireless networks, interference coordination techniques, radio resource scheduling, simulation processes, and protocols for radio access networks.

MEHRDAD SHARIAT received the B.Sc. degree in telecommunications engineering from the Iran University of Science and Technology, Iran, in 2005, and the Ph.D. degree in mobile communications from the University of Surrey, U.K., in 2010. He is currently a 5G Researcher with the Samsung R&D Institute UK (SRUK) contributing to 3GPP SA2 and 5G-PPP European projects on enabling network automation for 5G via a flexible architecture design. He has contributed to several U.K. and EC co-funded programmes as a Technical Lead and a Project Co-Investigator, including Mobile VCE (Core 4), FP7 (BeFEMTO, iJOIN, and MiWaveS), and 5G-PPP (mmMAGIC, METIS-II, and 5G-MoNArch).

MICHAEL EINHAUS received the Dipl.-Ing. and Dr.-Ing. degrees from RWTH Aachen University, in 2002 and 2008, respectively. Since 2015, he has been a Professor with the Leipzig University of Telecommunications (HfTL), where his research is focused on software defined radio, virtualization concepts, resource allocation and interference coordination in cellular networks, and radio channel modeling. Prior to his current occupation, he was for several years active in WiMAX Forum and 3GPP RAN1 standardization for NEC and Panasonic, respectively. He has authored numerous research publications and patents in the area of mobile radio communications.

QING WEI received the B.Sc. degree in electrical engineering from Shanghai Jiao Tong University, China, in 1997, and the M.Sc. degree in communication engineering from TU Munich, Germany, in 2001. She has been with DOCOMO Euro Labs in the area of mobile networking technologies, since 2002. In 2015, she joined Huawei GRC and works as a Principal Researcher in the area of 5G mobile network architecture. Besides numerous publications and IPRs, she has led and contributed to several EU projects including, End-to-end reconfigurability, WearIT, SASER-SaveNet, 5G-Xhaul, and 5G_MoNArch, as well as standardization activities, such as NGMN, 3GPP SA2/RAN3, and ETSI.

ÖMER BULAKCI received the B.Sc. degree from METU, Turkey, in 2006, the M.Sc. degree from TUM, Germany, in 2008, and the Ph.D. degree from Aalto University, Finland, in 2013. He was with Nokia Siemens Networks, Germany, on LTE-A relaying, from 2009 to 2012. Since 2012, he has been working toward 5G at Huawei GRC, Germany. He has authored/co-authored more than 80 publications, including patents and book editorship. He is leading the work package on overall architecture in 5G-MoNArch and is the 5GPPP Architecture WG Vice-Chairman.

ANTONIO DE DOMENICO received the M.Sc. degree in telecommunication engineering from the University of Rome âĂIJLa Sapienza,âĂİ in 2008, and the Ph.D. degree in telecommunication engineering from the University of Grenoble. Since 2009, he has been with CEA-LETI MINATEC, Grenoble, France, as a Research Engineer. He is the main inventor or co-inventor of 11 patents. His current research interests include cloud-enabled heterogeneous wireless networks, millimeter-wave-based communications, machine learning, and green communications.