Small cEllS coordinAtion for Multi-tenancy and Edge services

Grant Agreement No.671596

Topic: H2020-2014-ICT-14

Advanced 5G Network Infrastructure for the Future Internet

Research and Innovation Action

Deliverable D3.3

Framework of a distributed network management system capable to host and run Self-X functionalities

Document Number:	H2020-5GPPP-GA No.671596/WP3/D3.3/*30.06.2017*
Contractual Date of Delivery:	*30.06.2017*
Editor:	August Betzler (i2CAT)
Work-package:	*WP3*
Distribution / Type:	Public (PU) / Report (R)
Version:	1.0
Total Number of Pages:	96
File:	*SESAME_Deliverable 3.3_v1.0_Final*

Abstract

The SESAME architecture relies on a Self-X framework that introduces a large number of Self-X functionalities to assure a robust and highly efficient of the SESAME architecture. These Self-X features autonomously perform tasks such as the optimisation and maintenance of RAN and backhauling features and configurations, as well as the management and configuration of several elemental VNFs of the architecture. The Self-X features that compose the framework can be categorized into cSON, dSON, and hSON functionalities and further, each feature can be mapped to specific components of the SESAME architecture. Within the framework, the Self-X features and the SESAME components interact with each other, establishing the required control loops and workflows of the Self-X functionalities. Moreover, these interactions allow some of the Self-X features to consider input and to act beyond a local level, on an architectural-wide scope.

A general overview of SESAME's Self-X features has been provided in D3.2, introducing each feature separately. This deliverable extends D3.2 and puts the parts in a relation, providing a holistic view of SESAME's framework of a distributed network management system capable to host and run Self-X functionalities. Further, as part of an in depth technical description, this document provides detailed technical information on the elemental Self-X features that compose the framework, explaining how each feature is mapped to the SESAME architecture and how the features interact with the rest of the architecture.

Version History

Version	Date	Comments, Changes, Status	Authors, contributors, reviewers
0.1	04/05/2017	Initial Table of Contents	August Betzler (i2CAT)
0.1.1	15/05/2017	Updated Table of Contents	August Betzler (i2CAT)
0.1.2	25/05/2017	Updated Table of Contents	Jordi Pérez R. (UPC)
0.2	01/06/2017	Added contributions to Section 2	Contribution from several partners, merging: August Betzler (i2CAT)
0.3	13/06/2017	Added content to Section 1 and contributions to Section 2 from FBK-CNET	August Betzler(i2CAT) Leonardo Goratti (FBK-CNET) Shah Nawaz Khan Roberto Riggio
0.3.1	19/06/2017	Added further FBK-CNET contributions to Section 2	August Betzler (i2CAT) Leonardo Goratti (FBK-CNET) Shah Nawaz Khan (FBK-CNET) Roberto Riggio (FBK-CNET) Tejas Subramanya (FBK-CNET) Supreeth Herle (FBK-CNET)
0.4	19/06/2017	Swapped Section 2 and 3, added content to new Section 2	August Betzler (i2CAT)
0.5	23/06/2017	Contributions and partial review of Sections 2 and 3	Jordi Pérez R. (UPC) Alan Whitehead (IPA) Leonardo Goratti (FBK-CNET) Shah Nawaz Khan (FBK-CNET) Roberto Riggio (FBK-CNET) Tejas Subramanya (FBK-CNET) Supreeth Herle (FBK-CNET)
0.6	26/06/2017	Merging contributions, adding abstract, formatting, updating Glossary	
0.7	28/06/2017	Integrated changes and additions from IPA, UPC and CNET	August Betzler (i2CAT) Jordi Pérez R. (UPC) Alan Whitehead (IPA) Leonardo Goratti (FBK-CNET)
0.8	30/06/2017	Internal Revision of the document	Leonardo Goratti (FBK-CNET)
0.9	30/06/2017	Final check and corrections	August Betzler (i2CAT)
0.10	03/07/2017	Internal Revision of the document	Alexandros Kostopoulos (OTE)
0.11	04/07/2017	Internal Revision of the document	Athanassios Dardamanis (SMNET)
1.0	06/07/2017	Full conceptual and editorial review by OTE – Document ready for submission to the Commission	Ioannis Chochliouros (OTE)

Contributors

First Name	Last Name	Partner	Email
August	Betzler	i2CAT	*august.betzler@i2cat.net*
Daniel	Camps-Mur	i2CAT	*daniel.camps@i2cat.net*
Jordi	Pérez-Romero	UPC	*jorperez@tsc.upc.edu*
Oriol	Sallent	UPC	*sallent@tsc.upc.edu*
Ramon	Ferrús	UPC	*ferrus@tsc.upc.edu*
Ramon	Agustí	UPC	*ramon@tsc.upc.edu*
Juan	Sánchez-González	UPC	*juansanchez@tsc.upc.edu*
Ferran	Casadevall	UPC	*ferranc@tsc.upc.edu*
Alan	Whitehead	IP.Access	*alan.whitehead@ipaccess.com*
Leonardo	Goratti	CNET	*lgoratti@fbk.eu*
Shah	Nawaz Khan	CNET	*s.khan@fbk.eu*
Roberto	Riggio	CNET	*rriggio@fbk.eu*
Tejas	Subramanya	CNET	*t.subramanya@fbk.eu*
Supreeth	Herle	CNET	*s.herle@fbk.eu*
Athanassios	Dardamanis	SMNET	*adardamanis@smart.net.gr*
Alexandros	Kostopoulos	OTE	*alexkosto@oteresearch.gr*
Ioannis	Chochliouros	OTE	*ichochliouros@oteresearch.gr*

Glossary

Acronym	Explanation
3GPP	Third Generation Partnership Project
4G	Fourth Generation of Mobile Communications
5G	Fifth Generation of Mobile Communications
AB	Access Barring
AC	Admission Control
ACM	Association for Computer Machinery
AM	Acknowledged Mode
AMBR	Aggregated Maximum Bit Rate
ANR	Automatic Neighbour Relations
API	Application Programming Interface
ARFN	Absolute Radio Frequency Number
ARM	Advanced RISC Machine
ARP	Allocation and Retention Priority
ARQ	Automatic Repeat reQuest
AS	Access Stratum
AWGN	Additive White Gaussian Noise
BCCH	Broadcast Control Channel
BF	Broadband Forum
BS	Base Station
C-RNTI	Cell Radio Network Temporary Identity
CAC	Call Admission Control
CCCH	Common Control Channel
CESINR	Cell Edge SINR
CESC	Cloud-Enabled Small Cells
CESCM	CESC Manager
CMC	Common Messaging Calls
CMC	Connection Mobility Control
CN	Core Network
CP	Control Plane
CPE	Customer Premises Equipment
CPU	Central Processing Unit
CQI	Channel Quality Indicator
cSON	centralised SON
CWMP	CPE WAN Management Protocol
DB	Data Base
dB	decibel
DC	Data Centre
DL	Downlink
dSON	distributed SON
E2E	End-to-End
E-RAB	Evolved Radio Access Bearer
E-UTRAN	Evolved Universal Terrestrial Radio Access Network
EARFCN	EUTRA Absolute Radio-Frequency Channel
EMS	Element Management System
eMBB	enhanced Mobile Broadband

eNB	Enhanced Node B
EPC	Evolved Packet Core
ETSI	European Telecommunications Standards Institute
eV2X	enhanced Vehicle to Everything
FAP	Femto Access Point
FCAPS	Fault, Configuration, Accounting, Performance, Security
FFR	Fractional Frequency Reuse
FR	Frequency Reuse
GA	Grant Agreement
GAP	Generalized Assignment Problem
GBR	Guaranteed Bit Rate
GHz	Giga Hertz
GPRS	General Packet Radio Service
GS	Group Specification
GTP	GPRS Tunneling Protocol
GUI	Graphics User Interface
H2020	Horizon 2020
HARQ	Hybrid Automatic Repeat Request
hSON	hybrid SON
HTTP	HyperText Transfer Protocol
HW	Hardware
I/O, i/o	Input/Output
ICI	Inter-Cell Interference
ID, id	Identifier
IEEE	Institute of Electrical and Electronic Engineers
InH	Indoor Hotspot
ICT	Information and Communication Technology
IoT	Internet of Things
IP	Internet Protocol
ISD	Inter-Site Distance
ITU	International Telecommunication Union
KPI	Key Performance Indicator
LC	Load Control
LOS	Line of Sight
LTE	Long Term Evolution
MAC	Medium Access Control
MANO	Management and Organization
MBB	Mobile Broadband
MBR	Maximum Bit Rate
MC-GAP	Multi-Choice Generalized Assignment Problem
MC-KP	Multi-Choice Knapsack Problem
MCPTT	Mission Critical Push-To-Talk
MCS	Modulation and Coding Scheme
MEC	Mobuile Edge Computing
MEC	Multi-access Edge Computing
MHz	Mega Hertz
MKP	Multiple Knapsack Problem
MLB	Mobility Load Balancing
MME	Mobility Management Entity

mMTC	massive Machine Type Communications
MNO	Mobile Network Operator
MOCN	Multi Operator Core Network -
ms	millisecond
MTC	Machine Type Communications
NAS	Non-Access Stratum
NB	Narrow Band
NFV	Network Functions Virtualisation
NG	Next Generation
NGMN	Next Generation Mobile Networks
NIS	Network Information Service
NLOS	Non Line Of Sight
NMS	Network Management System
NP	Non-Polynomial
NR	New Radio
NWL	Network Listen
OAI	Open Air Interface
ODL	OpenDayLight
OF	OpenFlow
OFDM	Orthogonal Frequency-Division Multiplexing
OFDMA	Orthogonal Frequency Division Multiple Access
OTT	Over-The-Top
OvS	OpenvSwitch
PCCH	Paging Control Channel
PCI	Physical Cell Identity
PDCP	Packet Data Convergence Protocol
PHY	Physical Layer
PLMN	Public Land Mobile Network
PM	Performance Management
PNF	Physical Network Function
PoC	Proof of Concept
PPP	Poisson Point Process
PPP	Public Private Partnership
PRACH	Physical Random Access CHannel
PRB	Physical Resource Block
PS	Public Safety
QAM	Quadrature Amplitude Modulation
QCI	QoS Class Identifier
QoS	Quality of Service
QPSK	Quadrature Phase Shift Keying
RAB	Radio Access Bearer
RAC	Radio Admission Control
RAM	Random Access Memory
RAN	Radio Access Network
RAT	Radio Access Technology
RB	Radio Bearer
RB	Resource Block
RBC	Radio Bearer Control
REST	Representational State Transfer

RF	Radio Frequency
RIA	Research and Innovation Action
RISC	Reduced Instruction Set Computing
RLC	Radio Link Control
RNIS	Radio Network Information Service
RNTI	Radio Network Temporary Identifier
RON	Rise Over Neighbours
RRC	Radio Resource Control
RRM	Radio Resource Management
RSCP	Received Signal Code Power
RSP	Reference Signal Power
RSRP	Reference Signal Received Power
RSRQ	Reference Signal Received Quality
RSSI	Received Signal Strength Indication
RTT	Round-Trip Time
RX	Receiver
SAGBR	Scenario Aggregated Guaranteed Bit Rate
SC	Small Cell
SCNO	Small Cell Network Operator
SCTP	Stream Control Transmission Protocol
SDN	Software-Defined Network
SFC	Service Function Chaining
SE	Spectral Efficiency
SESAME	Small cEllS coordinAtion for Multi-tenancy and Edge services
SIB	System Information Block
SINR	Signal to Interference and Noise Ratio
SLA	Service Level Agreement
SON	Self-Organizing Network
SRB	Signalling RB
SW	Software
TAC	Tracking Area Code
TEID	Tunnel End-point Identifier
TM	Transparent Mode
TNPRE	Thermal Noise Per Resource Element
TOFS	Traffic Offload Function
TR	Technical Report
TS	Technical Specification
TTI	Transmission Time Interval
TX	Transmitter
UARFCN	UTRA Absolute Radio Frequency Channel Number
UE	User Equipment
UL	Uplink
UM	Unacknowledged Mode
uMTC	ultra-reliable MTC
UMTS	Universal Mobile Telecommunications System
UP	User Plane
URLLC	Ultra-Reliable and Low Latency Communications
UTRAN	Universal Terrestrial Radio Access Network
V2X	Vehicle to Everything

VIM	Virtualised Infrastructure Manager
VNF	Virtual Network Function
VNFM	Virtual Network Function Manager
VSC	Virtual Small Cell
VSCNO	Virtual Small Cell Network Operator
WAN	Wide Area Network
Wi-Fi	Wireless Fidelity
WP	Work Package
XML	eXtensible Markup Language

Table of Content

List of Figures

List of Tables

1 Introduction

SESAME relies on a set of core mechanisms that, *on one hand* assure the correct operation of the architecture and, *on the other*, optimize the efficiency and performance of the system. These Self-Organizing Network (SON) functionalities, also referred to as Self-X functions, a group of mostly autonomous mechanisms that must interact with the SESAME architecture. These mechanisms can be further divided into centralized (cSON), distributed (dSON) and hybrid (hSON) functionalities, depending on which resources they operate on, and where they are placed. The Element Management Systems (EMSs) and the Light DC are the main entities that host cSON features in SESAME. The dSON features are generally hosted in the Cloud Enabled Small Cells (CESCs). However, they can also be placed as part of the Light DC, as will be shown later. On the other hand, hSON features are hosted in both the CESCs and the EMSs 95[1].

Apart from determining the scope and domain in which the Self-X features are operating, they can further be categorized into Self-Planning, Self-Optimisation, and Self-Healing functions. Deliverable D3.2 gathers the proposals for Self-X functions that cover these three aspects of SON features in the SESAME architecture. The document studies each of the cSON, dSON, and hSON functionalities, revealing a large variety of mechanisms that are applied at different levels of the architecture, namely the CESCs (also the Light DC) and the CESCM.

The initial studies carried out in the previous SESAME Deliverable D3.2[1] serve as a "base" for the evaluations performed in this deliverable that presents the Self-X functions of SESAME "as a whole", a framework that is capable of hosting and running Self-X functionalities. This document is structured as follows: Section 2 describes the Self-X features of SESAME as a whole, providing a holistic view on the functionalities that form the Self-X framework and detailed discussions related to the RAN. Apart from that, further possibilities of configuring the framework externally are discussed. After this comprehensive introduction into the framework, an extensive description of the Self-X features that are going to be implemented in the SESAME architecture is given in Section 3. The section presents each of SESAME's Self-X functionalities from a technical point of view, also giving insights about their implementation. For each Self-X feature, it is explained how the Self-X feature "maps" to the SESAME architecture components. Eventually, a detailed description of each Self-X feature's workflow and its interactions with other Self-X features and SESAME components is provided. Section 4 concludes the deliverable with a summary of the content of this document.

[1] SESAME Deliverable D3.2: "*Self-X features and virtualised CESC multi-tenancy techniques evaluation*".

2 Framework for a Distributed Management Network

In deliverable D3.2 [1], a detailed overview of each Self-X feature present in the SESAME architecture was given. The large variety of Self-Planning, Self-Optimising and Self-Healing features presented in the deliverable "as a whole" forms what we call as the "Self-X framework" that ensures the correct and high-performance operation of specific key aspects of the SESAME architecture. Section 2.1 of the SESAME deliverable introduces the Self-X framework, giving a general view of all the Self-X features that form the SESAME's Self-X Framework, and where they are positioned in the related architecture. In Section 2.2, the scope focuses upon the specific aspect of Radio Access Network (RAN) slicing and radio interface protocol features, which can be considered a "core component" impacting on the Self-X framework and the way how Self-X functions can be developed. Finally, Section 2.3 offers a summary of configuration possibilities that are available to external entities to adjust the way the Self-X framework operates.

2.1 A Holistic Vision of the Self-X Framework in SESAME

The SESAME Self-X Framework is responsible for planning and optimising the configuration and operation of the RAN, the wireless backhaul and the VNF placement. Furthermore, it is responsible for the management of the Small Cells (SCs). The Self-X features provided in SESAME as a whole, deal with the assignment and management of radio resources in multi-tenant SESAME deployments, as well as the handling of VNFs. As a result of applying the Self-X functionalities, the system is capable of automatically picking the right network layout and radio frequencies for each of the SCs in a deployment. Also, the framework is capable of setting up the wireless backhaul and configuring the VNFs, while taking into account multi-tenancy requirements, such as the slicing of the RAN and backhaul, as agreed in the Service Level Agreement (SLA) of the tenants. The SLAs also determine how the Self-X framework behaves during operation: that is, the SCs perform automatic optimization of their radio settings and features for controlling the admission of new users, optimizing the handover procedure between SCs and maintaining the slicing in the RAN and wireless backhaul. These crucial tasks that require interactions between the Self-X features and the SESAME architecture are exposed in this document.

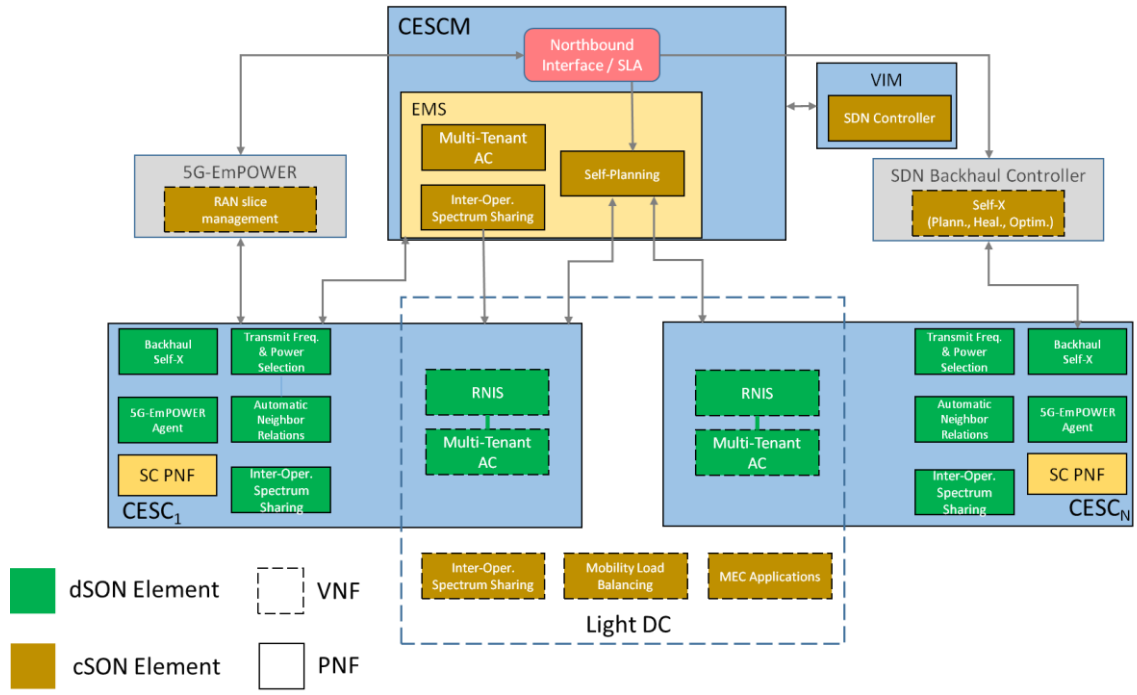

Figure 1: Overview of Self-X features that form SESAME's Self-X framework

Figure 1 shows all the Self-X features that "form" the Self-X framework in SESAME, where they are located in the architecture, and how they interact between each other and the architecture. The cSON features are classically associated to the management systems, i.e., the CESCM, where they run as PNFs (Physical Network Functions), having an overall view encompassing multiple CESCs. However, in SESAME, given the existence of the Light DC, it is possible to "place" as well some cSON functionalities at the Light DC, running as VNFs. Note that as depicted in this overview, some cSON features are not running in the CESCM, but as part of the SESAME architecture (e.g., the cSON features implemented in the 5G-EmPOWER or SDN Backhaul Controllers, which are introduced later). Like the Virtualised Infrastructure Manager (VIM), the 5G-EmPOWER controller responsible for managing the RAN slicing, and the SDN Backhaul Controller responsible for managing the wireless backhaul, are depicted to run outside of the CESCM. The cSON features, in general, maintain an overview and control over a specific subset of features, such as the RAN or the wireless backhaul. For that, basic information is exchanged between the cSON features and the different components of the SESAME architecture: information available to the EMS is consumed by several cSON features, like the multi-tenant access control. In this case, the cSON features are even collocated in the EMS, as they form part of it.

The dSON features depicted in Figure 1 are a mix of VNFs and Physical Network Function(s) (PNF(s)) that run on the CESCs, either as part of the Light DC or as part of the SC, *respectively*. Basically, all dSON features in the framework are responsible for configuring, maintaining and optimising the use of the access or backhaul radio elements. This process often requires the exchange of data between the cSON and dSON components and sometimes even external input, e.g., the requirements on the network slicing prescribed by the SLAs.

The Self-X features that form the Self-X framework have different degrees of autonomy. Fully autonomous features do not require any external input, i.e., once they are running they automatically configure internal parameters and they also communicate autonomously with other SESAME components to obtain required information or to store information. An example

for an autonomous feature that operates and "adjusts" on its own is the autonomous scan of the radio environment and the subsequent adaptation of operating radio frequencies (RFs) to be used by the SC. Autonomously running features also might require data that can be obtained from other components within the SESAME architecture, i.e., from components such as the EMS that collects data which can be used to apply Self-X features located in the CESCM, the Light DC or in the SCs. This mainly affects cSON features. In general, this type of interactions is quite basic: a Self-X feature can use architecture-internal APIs to request information that is required to operate from a component. The EMS provides the correspondent interfaces to request the data. On the other hand, if a Self-X feature generates data that can be shared internally within the architecture, it stores this information either locally or in a central access point, as it would be the CESCM. For these basic interactions, the different SESAME components use simple APIs that are based on well-known protocols, such as Representational State Transfer[2] (REST), or they directly write/read from shared files that are accessible to all or a subset of Self-X features. The simplicity of the communication and access model between the Self-X components and the architecture makes it a very flexible model that can be extended easily. The use of standard protocols ensures the interoperability between the Self-X components and the rest of the architecture. Eventually, it also should be stated that within the group of fully autonomous features the exchange of basic information is designed in a way that it can be performed automatically, as part of the operations a Self-X feature carries out during runtime.

On the other hand, there are Self-X features that require an external configuration in order to initiate/operate correctly -or to deliver- an optimal performance. Such an input can, *for example*, come directly from a Virtual Small Cell Network Operator (VSCNO) in order to adjust a specific parameter of the feature. This can be, *for example*, a configuration parameter that affects the behaviour of a VNF assigned to the VSCNO in a deployment. In this case, the configuration is performed directly by the VSCNO, via a dedicated API. Otherwise, the configuration of a Self-X feature can be performed by the CESCM or by one of its sub-components: either information stored in the EMS is retrieved by the Self-X feature or the CESCM "translates" any type of requirements from the VSCNOs (i.e., as determined in the SLAs) into data that can be used by the Self-X feature. An example for such a case is the translation of an SLA that assigns a set of CESCs to an API call that can be interpreted by the wireless backhaul SDN controller: the SDN controller needs to be instructed about which CESCs will be available to the tenant, in order to configure the devices in the wireless backhaul. Continuing with this example, in the next step of the backhaul configuration, the CESCM needs to "tell" the backhaul controller, which airtime ratio will be applied on the tenants' slice of the wireless backhaul. These requirements are SLA requirements translated to a RESTful API call that is understood by the backhaul controller. The controller then translates these calls to internal API calls (for example, the configuration of the backhaul nodes via NETCONF). After receiving the configuration call, the dSON features running in the CESCs have all the information necessary to operate correctly. This chain of interactions (VSCNO -> SLA -> CESCM -> cSON -> CESC -> dSON) shows how information enters the framework, how it is translated into a format that is understood by the internal components and handed over between SESAME components and the Self-X features via the internal framework APIs.

[2] Representational state transfer (REST) or RESTful web services is a way of providing interoperability between computer systems on the Internet. REST-compliant Web services allow requesting systems to access and manipulate textual representations of Web resources using a uniform and predefined set of stateless operations. For more related information see, *inter-alia*: https://en.wikipedia.org/wiki/Representational_state_transfer

The following paragraphs give a comprehensive overview of the Self-X features that can be found in the SESAME architecture and their basic interactions. The details on a *per feature* basis about the specific interactions between Self-X features and the SESAME architecture are provided in Section 3.

The increasing softwarization of the network and autonomous behaviour to carry out several operations enable efficient management of both network and radio resources. Exploiting the NFV MANO [2] management framework in the context of SESAME (in the Light DC virtualisation environment), it is possible to create, delete and modify network functions that are implemented as VNF, as well as create, delete and modify composite network services. Both Self-X and the NFV MANO stem as the constituents of a powerful cognitive management framework. One key feature that is native in the SESAME system is resource sharing: the RAN, as well as compute and storage that are localised at the edge of the network. In this regard, it is also worth noticing that the SESAME system is well positioned with respect to the paradigm of multi-access edge computing (MEC) [3]. The edge-cloud resources are usually more limited if compared to a large data centre facility but the edge-cloud has the advantage of keeping more localised the traffic created by the users that make access to different services. Furthermore, it offers a stimulating environment where to develop new features that use efficiently the edge-cloud environment. In the context of SESAME, a few Self-X features will be presented in sections 3.1, Section 3.2 and Section 3.3. More in specific, Section 3.1 resumes and brings to conclusion the work already started in [1] regarding the proposed spatial resource scheduler. This is a way of Self-Optimising the performance of a CESC, a group of CESCs or the whole CESC cluster with respect to the problems of Inter-Cell Interference (ICI) and the selection of the best feasible Modulation and Coding Scheme (MCS) for packets that have to be transmitted in downlink. Depending on whether the spatial scheduler runs, a cSON or a dSON can be realised. Results for the proposed technique will be shown through Matlab simulations. The work presented in Section 3.2, describes RAN sharing and the implementation of this feature which, *as discussed above*, is key in SESAME. RAN sharing is indeed crucial to enable customised slicing of resources that can be autonomously reconfigured in a Self-Organising manner as the load in tenant networks varies or interference degrades the performance. The design principles of the RAN sharing function, together with results related to the implementation will be provided. Finally, Section 3.3 provides the description of the feature of edge content caching and its implementation. Caching of content is meant to spare backhaul resources and offer services with reduced latency to the end users, which is particularly crucial for accessing video contents. In the context of SESAME this can be done in an autonomous manner through a Self-Organising network feature. In case of video for instance, a specific content can be buffered based on a popularity index and missed cache rate as studied in [4] and in [5]. Also in this case, the design principles of the edge content caching feature and the results related to its implementation will be presented.

The Self-Planning functionality of the PNF associated with a CESC is a dSON feature that operates within bounds defined by the EMS. Transmit parameter selection allows the PNF to set its operational parameters in order to minimise interference both from and to surrounding cells and to achieve a coverage target specified by the EMS. An in-depth description of this feature is provided in section 3.4.

Automatic Neighbour Relations (ANR) is a dSON feature that allows the PNF to detect and select neighbour cells that are suitable targets for handout. In the SESAME multi-operator environment, the selection of candidates has to consider both all of the PLMNs supported by the PNF and all of the PLMNs supported by each detected neighbour. In addition, as space in a

physical cell's neighbour lists is constrained, the PNF has to ensure that each of the PLMNs that it hosts are given fair representation with respect to the proportion of configured neighbours that support each PLMN. See section 3.5 for a more detail discussion of these issues.

The Self-Optimisation of multi-tenant Admission Control (AC) intends to adjust the AC thresholds for admitting/rejecting the Evolved Radio Access Bearer (E-RAB) establishments of the different tenants sharing the CESCs. The Self-X function is a hybrid SON approach where the dSON performs the adjustment based on the local information about the resource occupation of the different tenants in the CESC, while the cSON takes into account the overall resource consumption considering all the CESCs in the scenario as a whole. The dSON and the AC are executed at the Light DC, while the cSON component is executed at the EMS. The details about these functionalities, including the considerations about implementation, are given in Section 3.6.

The challenge of providing multi-tenancy connectivity between SCs and the tenant's core networks in a deployment is handled in an efficient and flexible way via the SDN-*based* wireless backhaul. A series of Self-X features are responsible for Self-Planning, Self-Optimising and Self-Healing of the wireless backhaul. The backhaul' s Self-X features can be split in two main parts, the cSON part which is implemented in the SDN controller and the dSON part, which runs in the SCs. As detailed above in an example for a chain of interactions in the SESAME architecture, in order for the Self-X features to operate correctly, the exchange of information between the SESAME architecture and the features that form the wireless backhaul is necessary. In particular, the SDN controller requires input from the SLA management entity, in order to know which SCs need to be assigned to a tenant. This is a basic requirement, as it determines which of the physical devices that are the SCs form part of a tenant' s backhaul network, which requires to set up and configure the wireless interfaces of the SCs. Further, information needs to be handed over to the controller in order to assign each tenant's airtime ratio of their virtual network slice. In order for the wireless backhaul to provide a high and steady capacity of the wireless links that carry the end-to-end traffic between a CESC and the evolved packet core (EPC) network, a careful frequency selection for the wireless links of the backhaul is required. The planning needs to take into account possible self-interference of the backhaul radio, but also interference from external sources. Since the wireless backhaul operates in the unlicensed sub 6 GHz band, possible interference issues with RAN devices are not given. On the other hand, there is currently no mechanism implemented that adjusts the chosen frequencies to interferences from outside of the SESAME architecture (e.g., public Wi-Fi networks that operate close or along the SESAME deployment). As such, the wireless backhaul performs the wireless link channel assignment completely on its own as part of the Self-Planning feature, not requiring any additional input from any external component. Section 3.7 explains the cSON Self-Planning feature that is responsible for setting up and configuring the wireless backhaul. Section 3.8 introduces the Self-Optimisation features that are active during runtime and that are present both as cSON and dSON features. To conclude, the in technical description of the Self-Healing features in SESAME's Self-X framework are introduced in Section 3.9.

The Self-Planning functionality for multi-tenant SC networks intends to determine the updates that have to be performed in the network (e.g., addition of new CESCs, modification of the spectrum assignment, etc.) in response to specific situations such as the addition of a new tenant in the scenario, the modification of the SLAs of some tenants, etc. It is a cSON function that will be associated to the EMS or the NMS. The details about this function are presented in section 3.10.

The holistic view given in this subsection on the Self-X framework encompasses the key Self-X features that give support to and carry the main operations of the SESAME architecture. Being a crucial part of the architecture, we dedicate an additional subsection to the discussion of the specific role of the radio interface protocol features and the configuration for the RAN slicing in SESAME.

2.2 Radio Interface Protocol Features and Configuration for RAN slicing

Unlike current 4G systems mainly designed to provide a "one size fits all" solution for mobile broadband connectivity services, 5G systems are expected to be able to simultaneously support a wider range of application scenarios and business models resulting from the anticipated pervasive adoption of 5G technologies across different market segments (e.g., automotive, e-health, utilities, smart cities, agriculture, media and entertainment, high-tech manufacturing) and the introduction of more flexible and agile service delivery models (e.g., Network as a Service, 3rd party customization through network capabilities exposure) [6]. This expected versatility comes with a high variety of requirements on network functionalities (in terms of e.g., security, mobility, charging, policy control) and expected performance (e.g., target peak data rates of 20Gbps / 10Gbps for downlink / uplink, latencies below 1 ms with reliability levels of about 10^{-5}, mobility target of 500 km/h, three-fold increase in spectrum efficiency) that cannot always be "met" at the same time through a common network setting (e.g., optimizing the network for low latency with high reliability could come at the expense of reduced spectral efficiency). In this context, support for network slicing in 5G systems has been established as a foundational requirement to allow 5G system operators to compose and manage dedicated logical networks with specific functionality, without losing the economies of scale of a common infrastructure ([7], [8]). Each one of these logical networks, referred to as *network slice*, can be tailored to fulfil at least a couple of purposes:

- To provide a particular system behaviour (i.e., slice type) through the use of specific control plane and/or user plane functions to best support specific service / applications domains (e.g., optimized protocols for enhanced mobile broadband (eMBB), massive Machine Type Communications (mMTC), Ultra-Reliable and Low Latency Communications (URLLC), enhanced Vehicle to Everything (eV2X)). For instance, a User Equipment (UE) for smart metering applications can be served through a network slice with radio access tailored to very small, infrequent messages and with no need to implement unnecessary functions and features (e.g., no mobility support, no carrier aggregation features, simplified QoS framework).

- To provide a particular tenant (i.e., an organization or business entity entitled to use the network slice) with a given level of guaranteed network resources and isolation with regard to the operation of other concurrent network slices. For instance, a tenant can be a Public Safety (PS) agency whose UEs/subscribers can be served through a network slice that guarantees a minimum capacity for eMBB access even during network congestion periods.

3GPP has already completed the normative specifications regarding service and operational requirements in order to support network slicing [9]. Initial work has also started at system architecture level with the main focus placed on the required functionality for network slice identification and selection [10]. Simultaneously, the network slicing concept is being addressed

in the 5G architectures currently under development in different research projects such as 5G-NORMA [11], METIS-II [12] or SESAME [13]. However, while important progress exists on the realization of network slicing within the Core Network (CN) functionality of 5G systems, greatly facilitated by the applicability of Network Function Virtualization (NFV) and Software Defined Networking (SDN) technologies, slicing the RAN part still poses multiple open questions.

In this context, this section firstly tackles the overall network slicing architectural framework for 5G systems, and clearly identifies the fundamental design challenges for the realization of the RAN slicing. Then, a solution approach addressing the open questions is proposed through the definition of a set of descriptors to parametrize the features, policies and resources in place across the radio protocol layers for the configuration of a RAN slice.

The systematic and comprehensive analysis conducted in this section intends to tangibly structure and specify how a RAN slice can be realized and how various RAN slices can be separately customized. In this way, the RAN slicing framework presented here enables the separate optimization of each slice by means of the appropriate Self-X functions. Finally, the applicability of the proposed RAN slicing solution framework is illustrated through different RAN deployment scenarios involving multiple service providers, service types and groups of subscribers/applications.

2.2.1 Architectural Framework and RAN Slicing Design Challenges

The realization of network slices within the Next Generation (NG) architecture for 5G systems considers, in the most general case, support for network slicing within both the CN and RAN functionality. The NG RAN consists of a number of nodes that provide the radio protocol termination towards the UEs. A NG RAN node, which can support one or multiple cells, is denoted as gNB if its radio interface is based on the New Radio (NR) specifications and as eLTE eNB if the interface is based on the evolved LTE specifications [14]. The RAN nodes are interconnected with each other by means of the Xn interface and with the NG CN through the NG2 and NG3 interfaces for the control plane (CP) and user plane (UP), *respectively*. As illustrated in Figure 2, a quite diverse range of UEs (e.g., smartphones, professional video cameras, PS terminals, Internet of Things (IoT) devices, car UE terminals) can be connected to the same NG RAN node, though potentially served via different network slices. A network slice is realized by involving a particular combination of RAN functions (referred to as the RAN part of the network slice or simply RAN slice) and CN functions (referred to as the CN part of the network slice or simply CN slice). Within a mobile network identified by a Public Land Mobile Network (PLMN) identity (ID), each network slice is to be uniquely identified by a *Slice ID*, which could take standard values to facilitate slicing configurations across networks in roaming scenarios or just remain PLMN-specific [10].

As illustrated in Figure 2, the CN slices are to be realized through a set of common CP network functions (e.g., subscriber authentication and authorization) and a set of slice-specific CP and UP network functions (e.g., dedicated mobility and session management, dedicated encapsulation and tunnelling). All these network functions are likely to be implemented as virtualized network functions (VNFs) running on cloud infrastructures and flexibly orchestrated as required. On this basis, the accomplishment of different system behaviours in the CN part is based on the deployment of the required slice-specific CN functions while the necessary level of guaranteed network resources and isolation between concurrent CN slices can be well sustained in the use of multi-tenant cloud infrastructures supporting performance management and assurance [15].

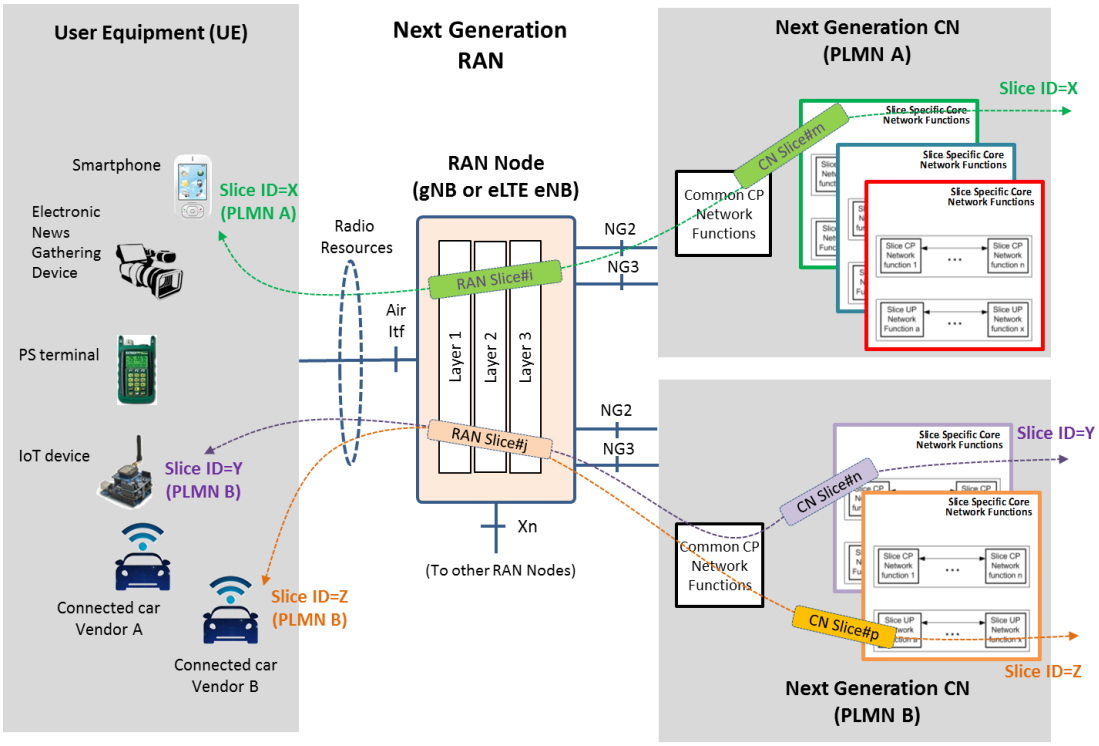

Figure 2: Illustration of the network slicing architecture

On the other hand, the realization of the RAN slices requires addressing how the pool of radio resources (i.e., RF bandwidth) allocated to a RAN node can be operated to simultaneously deliver multiple and diverse RAN behaviours, turning the RAN slicing support into a much more challenging issue. In this respect, let us consider that the RF bandwidth available for the operation of the NG RAN node can be flexibly arranged into a number of RF channels with diverse transmission bandwidths. A simple approach to support multiple RAN slices with slice-specific/optimized waveforms would be to implement each slice type on different cells working on separate RF channels and with separate radio protocol layer instances (i.e., Layer 1, 2 and 3 functions), in a similar way as it is done today with the multiple Radio Access Technologies (RATs) that co-exist in the current RAN infrastructures (e.g., UMTS cells, LTE cells, NB-IoT cells). However, scalability in terms of the number of supported slices types and reduced resource efficiency (i.e., no traffic multiplexing gains among the RAN slices) would be clearly major drawbacks. Therefore, the fundamental design challenge strives in the realization of multiple RAN slices that can be concurrently multiplexed in a single cell while at the same time achieving an efficient use of the radio resources. This entails tackling the following open questions:

- How to arrange a number of slice-specific radio interface protocols over the same cell.
- How to manage the radio resource allocation to UEs within a cell so that some established levels on capacity and isolation can be offered per RAN slice.
- How to enable support for optimized Radio Resource Management (RRM) configurations and policies on a per-slice basis (e.g., admission policies, mobility control) and for supporting per-slice Self-X mechanisms for configuring the RRM of each slice.

2.2.2 Realization of RAN slices

Taking as starting point the radio interface protocol architecture in current 3G/4G RANs and the initial work conducted within 3GPP with regard to the NG RAN [14][16], the analysis presented in this section shows that a set of new blocks of information, configuration descriptors and protocol features has to be introduced across the protocol layers of a NG RAN node in order to address the mentioned open questions. In this respect, the proposed overall framework for RAN slicing support within a NG RAN node is illustrated in Figure 3 and explained in the following.

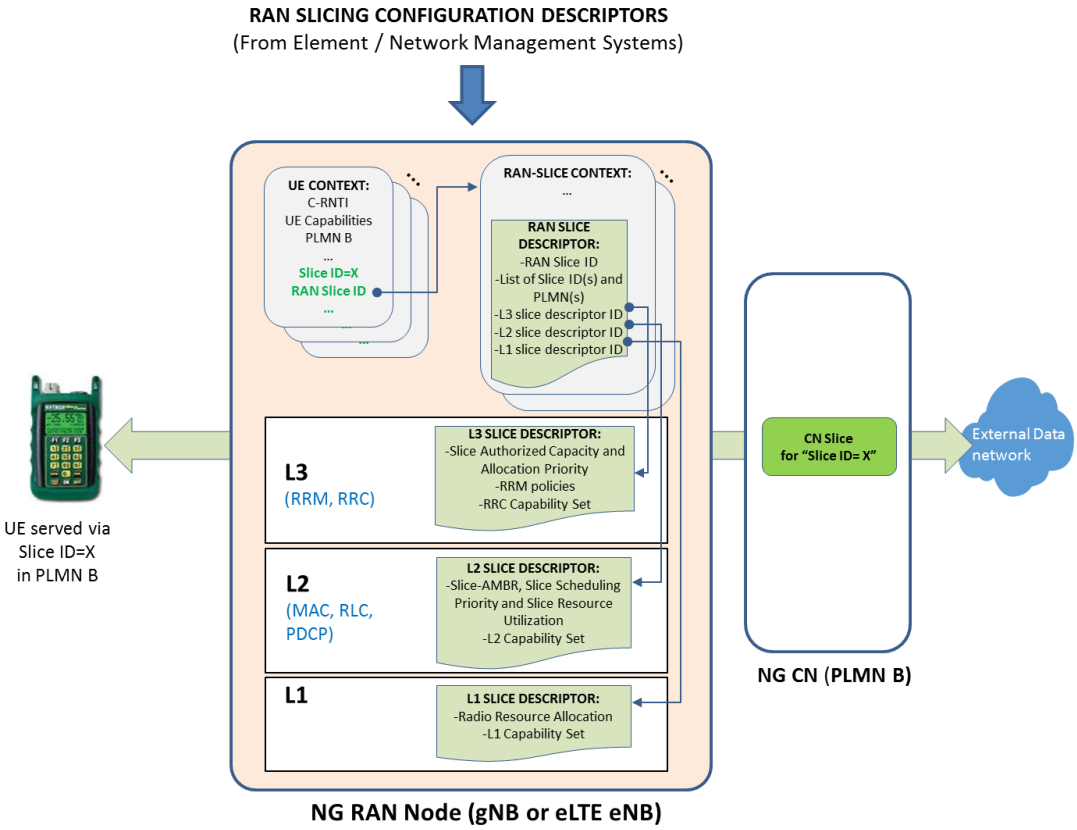

Figure 3: Framework for the realization of RAN slices in a NG RAN node

From an operational perspective, for a UE to get the services of a legacy 3G/4G RAN, a *UE Context* is instantiated within the RAN at the time that the UE becomes active (e.g., the UE is switched on or a service request is triggered) and a logical control connection (i.e., Radio Resource Control (RRC) connection) is established. The *UE Context*, which is populated from information exchanged between the RAN and both the UE and CN, is a block of information that contains all the necessary data required to maintain the RAN services towards the UE (e.g., temporary identifiers of the UE within the RAN such as the Cell Radio Network Temporary Identity [C-RNTI], security context, roaming and access restrictions, UE capabilities, QoS related information, etc.). On this basis, we propose to add a RAN slice identifier (*RAN Slice ID*) to the *UE Context* and use it as a pointer to a new block of information within the NG RAN node, denoted as *RAN Slice Context*, containing all the data necessary to support the operation of a particular RAN slice along with the *Slice ID(s)* that are served through the RAN slice. The selection of the *RAN Slice ID(s)* is to be conducted by the RAN node based on the *Slice ID(s)* signalled by the UEs during the initial attach procedure or indicated from CN in subsequent signalling (the specific signalling procedures are still work in process within the 3GPP [10]).

Accordingly, from a functional perspective, it is proposed to specify the operation of each RAN slice through a set of configuration descriptors used to parametrize the features, policies and resources that have to be put in place across the radio protocol layers of the RAN node for the realization of the RAN slice. Particularly, a *RAN Slice Descriptor* is introduced as the baseline descriptor for the instantiation of a *RAN Slice Context* within a RAN node. Together with the *RAN Slice ID*, the *Slice Type* (e.g., eMBB, mMTC, etc.) and the list of associated *Slice ID(s)* and *PLMN ID(s)*, as illustrated inFigure 3, the *RAN Slice Descriptor* includes the pointers to the configuration descriptors of the underlying radio protocol layers 3, 2 and 1 (L3, L2, L1) for the realization of the RAN slice:

- *L3 slice descriptor*. L3 comprises the control functionality needed for the activation and maintenance of the Radio Bearers (RB), which are the data transfer services delivered by the radio protocol stack between the RAN and UE. This control functionality is realized by means of the RRC protocol for signalling exchange and by a number of RRM functions such as Radio Bearer Control (RBC) to set up, configure and release the RBs, Radio Admission Control (RAC) to check if there are enough resources for RB establishment and Connection Mobility Control (CMC) to guide the cell (re-)selection process by UE terminals and perform handover of the established RBs between cells as necessary. Therefore, a *L3 slice descriptor* is necessary to specify the capacity allocation for the RAN slice (e.g., number and characteristics of the RBs that can be simultaneously established, allocation priority for conflict resolution across slices), the RRM policies that govern the operation of the slice (e.g., RB configuration policies, mobility policies, congestion policies) and the capability set of the RRC protocol in use (e.g., application type specific RRC messages).

- *L2 slice descriptor*. L2 comprises a Medium Access Control (MAC) sub-layer for multiplexing and scheduling the packet transmissions of all the established RBs over a set of transport channels exposed by L1. Moreover, L2 embeds a number of processing functions configurable on a per-RB basis for e.g., segmentation, Automatic Repeat reQuest (ARQ) retransmissions, compression and ciphering (i.e., Radio Link Control (RLC) and Packet Data Convergence Protocol (PDCP) sub-layers in 3G/4G RAN). Therefore, considering that the current operation of the MAC sub-layer is based on individual UE and RB -specific QoS profiles, a *L2 slice descriptor* is necessary to be able to define packet scheduling behaviors that can be enforced on the traffic aggregate of RBs associated with the same slice and to specify the capability set of the applicable L2 processing functions.

- *L1 slice descriptor*. L1 provides L2 with transfer services in the form of transport channels, which define how the data is transferred (e.g., transport block sizes, Transmission Time Interval (TTI), channel coding, interleaving, rate matching). L1 also establishes the corresponding radio resource structure of the cell radio resources (e.g., waveform characteristics and time- / frequency- domain resource structure). Considering that a RAN slice may require specific L1 transfer service capabilities (e.g., low latency shared transport channel) and / or specific radio resource allocation of the cell radio resources, a *L1 slice descriptor* is needed to specify both aspects.

More details on the above introduced descriptors and their impact on the configuration and extended features necessary for RAN slicing within L3, L2 and L1 are discussed in the following.

2.2.2.1 L3 configuration

When multiple RAN slices are realized over shared radio resources, the RRM functions for RBC, RAC and CMC have to assure that each RAN slice is getting the expected amount of resources and, in case, handle any resource conflicts that might appear across slices. These RRM functions are typically implemented as vendor-specific algorithms only abided by the use of the data models and control signalling capabilities established in the standards (e.g., UE and RB QoS model parameters, UE Capabilities, RRC protocol messages and procedures). Therefore, we propose to specify the following set of parameters per RAN slice to dictate the operation of the RRM functions for capacity allocation and traffic isolation among the slices:

- *Slice Authorized Capacity*. This can be a combination of resource-*oriented* and rate-*oriented* parameters that limit the number and characteristics of the RBs established for the entire slice. Resource-*oriented* parameters can include absolute or relative occupation levels of the radio resources consumed by the RAN slice (i.e., radio resource limitations) as well as hard limits on the number of *UE Contexts* / RBs that can be simultaneously established (i.e., license limitations). Rate-*oriented* parameters can include rate limits on the aggregate bit rate of the entire set of admitted Guaranteed Bit Rate (GBR) RBs within the slice. All these slice capacity parameters are to be used by the RAC for the admission/rejection of RBs as well as by the RBC/CMC functions in order to decide on modifications, handovers or even forced releases of active RBs if network dynamics turns into increased radio resource consumption in excess of the established limits.

- *Slice Allocation Priority*. This parameter allows for conflict resolution amongst UE/RB resource requirements across slices that cannot be solved based only on the *Slice Authorized Capacity* parameters nor using the policies associated with the individual UE/RBs (e.g., a situation in which there are two GBR RB admission requests with the same UE/RB QoS profile but associated with distinct RAN slices that cannot be granted simultaneously due to temporary congestion). In legacy 3G/4G RANs, priority and pre-emption policies at UE/RB level are solved through the so-called Allocation and Retention Priority (ARP) parameter included in the UE/RB QoS profiles. The ARP encodes information about priority level (scalar with 15 levels), pre-emption capability (flag with “yes” or “no”) and pre-emption vulnerability (flag “yes” or “no”). On this basis, a similar semantic can be adopted for the *Slice Allocation Priority* parameter so that it could be used to override the UE/RB ARPs for congestion conflict resolution among slices.

Moreover, since pursuing different optimization targets for the efficient use of the radio resources requires RRM functions with different parameterization (e.g., different handover margins per RAN slice), the RRM policies should be specifiable on a per-slice basis. This can be accomplished following two different approaches:

- A prescriptive specification, establishing the RRM algorithms and associated configuration parameters (e.g., thresholds, timers) to be used. For instance, a scenario where the RAN is deployed by a neutral hosting provider and the Mobile Network Operator (MNO) who contracts the RAN slice is interested in keeping a tight control of the RRM operation of the RAN slice. This facilitates that the MNO can implement its own Self-X functions for configuring the RRM algorithms in its slice.

- A declarative specification, establishing the expected RRM behaviour (e.g., Key Performance Indicators (KPIs) and optimization goals). This approach could be the preferred solution in scenarios where the service provider who contracts the RAN slice does not matter on the implementation details but just expects a pre-established behaviour (i.e., the descriptor strictly defines what is wanted and the system decides how it is implemented). In this case, the Self-X functions would consider as input the expected RRM behavior and would translate it into a configuration of the RRM parameters for each slice.

With regard to the RRC protocol, the customization of RRC messages and procedures on a per-slice basis (e.g., specific RRC signalling for Narrow Band – Internet of Things (NB-IoT) applications) can be realized without any impact on the protocol stack if the customized RRC signalling is only used through UE-dedicated Signalling RBs (SRBs) and/or through common logical channels not shared with other RAN slices. However, if multiple RAN slices are configured to share the same set of common logical channels (e.g., Broadcast Control Channel (BCCH), Paging Control Channel (PCCH) and Common Control Channels (CCCHs)), the following extended features have to be incorporated within the RRC protocol to properly support RAN slicing:

- Protocol fields within the RRC messages to allow UEs to discriminate among signalling belonging to different slices.
- System Information Block (SIB) messages to advertise through a common BCCH the *Slice ID(s)* that can be reached from the cell. This allows the UE to take into account this information for network discovery and selection processes.
- SIB messages to support cell (re-)selection parameters and neighbouring cell information broadcasting per *Slice ID*, so that the behaviour of terminals in idle mode can be set differently for different slices.
- SIB messages to support Access Barring and Load Control (AB / LC) per *Slice ID* so that un-scheduled transmissions over the uplink CCCH can be controlled separately per slice. In this way, UE attempts to transmit can be prioritized to prevent that an overload situation in one slice could hinder the operation of other slices or even cause loss of service. That may be accomplished comparable to today's separated AB / LC that is provided per PLMN operator for network sharing.

Paging configuration features allowing paging cycles to be organized considering the specifics needs of each *Slice ID*.

2.2.2.2 L2 configuration

In current 3G/4G systems, radio resource scheduling by the MAC takes account of the traffic volume and the radio conditions per UE together with a set of QoS parameters associated with each UE and its established RBs that defines the expected forwarding behaviour per UE/RB (e.g., parameters such as QoS Class Identifier (QCI), Guaranteed Bit Rate (GBR), Maximum Bit Rate (MBR), Aggregated MBR per UE (UE-AMBR)). While this approach is sufficient to provide QoS differentiation on the packet forwarding treatment per-UE/RB, it lacks of any formalization to parameterize the expected QoS behaviour at packet forwarding level for the traffic aggregate of the group of UE/RBs associated with a given RAN slice.

Therefore, the proposed *L2 slice descriptor* includes the following set parameters to dictate, in combination with the per-UE/RB QoS profiles, the operation of the MAC scheduler and yield traffic isolation at packet forwarding level on a per-slice basis:

- Slice-AMBR, to limit the aggregate bit rate of all the Non-GBR RBs associated with the slice.
- Slice Scheduling Priority, to handle short-term traffic congestion conflicts between RBs with the same QoS profile (e.g., same QCI) but belonging to slices that should be given different precedence treatment.
- Slice Resource Utilization, used to establish constraints on the amount of physical layer resources scheduled by the MAC that are consumed by the slice. This constraint can be formulated as a percentage of the overall L1 resources that are managed by the MAC scheduler.

In addition, the *L2 slice descriptor* includes a *L2 capability set* to establish the possible configuration options for the RBs associated with the slice, including Hybrid ARQ configurations (e.g., synchronous / asynchronous operation and number of processes in parallel), RLC operation modes (e.g., Acknowledged / Unacknowledged / Transparent modes [AM / UM / TM] and status reporting) and PDCP options (e.g., ciphering support).

2.2.2.3 L1 configuration

The new physical layer in 5G (i.e., NR) is being defined with the target to provide high flexibility for the use of different waveforms (e.g., orthogonal frequency-division multiplexing (OFDM)-*based* waveforms with different numerologies for subcarrier spacing and cyclic prefix extensions) and adaptable time-frequency frame structures (e.g., selectable slot durations, support for mini-slots with a length as short as a single OFDM symbol, dynamic assignment of DL/UL transmission direction). Considering that the L1 optimal settings can differ per slice type, the proposed L1 descriptor intends to establish a partitioning of the L1 radio resource structure so that different L1 optimization settings can be simultaneously applied.

The proposed approach is illustrated in Figure 4. Each tile represents the smallest allocation unit of time duration Δt and frequency size Δf (e.g., Δt and Δf could be an integer number of, respectively, OFDM symbols and subcarriers with the minimum supported subcarrier separation). On this basis, the overall RF carrier resources are split-off in a number of *L1 slices* that could be separately optimized to offer specific transfer service capabilities (e.g., different TTI sizes with a minimum value and resolution given by Δt, synchronous/asynchronous access). The mixing of L1 slices in the same resource grid with different waveforms characteristics could be achieved through the use of OFDM-waveforms with scalable numerologies and transmitter windowing techniques that mitigate the potential inter-slice interference [17]. Moreover, this segmentation of resources per slice avoids the transmission of common reference and control signals over the entire RF bandwidth but instead use self-contained transmissions per slice that can be arranged as needed (e.g., one slice could be configured to use a strict time-division separation of the physical layer control and data, like in current LTE, while in another slice control and data could be multiplexed over the same radio resources [18]). Note that this approach requires some minimum amount of resources to be allocated within the resource grid structure to facilitate the cell search process (e.g., common reference signals for UEs to acquire time and frequency synchronization) and provide the information for UEs to locate the control channels within the L1 slices. This approach also allows the segmentation of resources to be

dynamic (e.g., the location and number of resource units allocated to each L1 slice adapted over time to match traffic demand variations).

An independent scheduler that spans across L1 and L2 slicing and where each slice has its own independent scheduler and resource blocks are assigned to different tenants in non-overlapping manner is presented in Section 3.2.

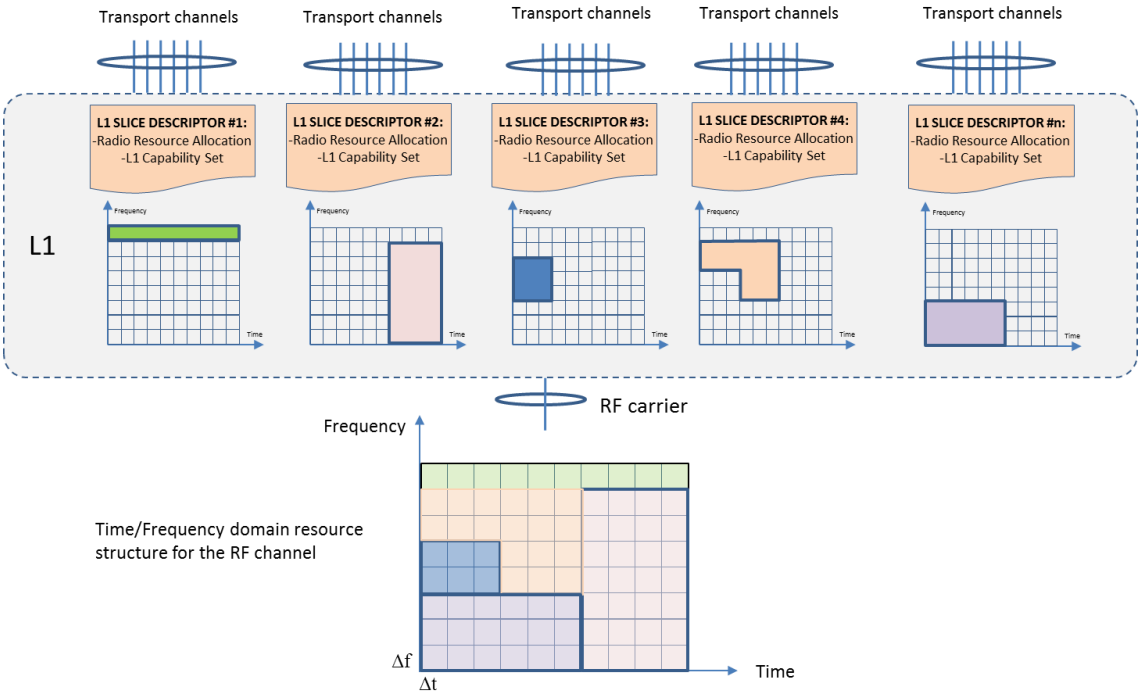

Figure 4: Illustration of slicing support in L1

2.2.3 RAN Slicing configuration examples

In order to gain insight into the proposed framework, Table 1 describes three illustrative scenarios for the applicability of RAN slicing, including the cases of: (1) a RAN owned by a commercial MNO; (2) a RAN deployed by a PS communications provider, and; (3) a neutral host provider deployment. For each scenario, a discussion on the business considerations that motivate the use of RAN slicing and a plausible RAN slicing configuration is provided. Additionally, the first case is developed in deeper detail, visualizing the radio interface protocol architecture along with specific parameters involved in the RAN slicing configuration descriptors.

Table 1: Illustrative scenarios for the applicability of RAN slicing

RAN deployed by a commercial MNO
A commercial MNO (MNO#A - PLMN#A) deploys a RAN and primarily exploits it to deliver MBB and IoT services. To this end, MNO#A sets up three RAN slices: one for eMBB and two for mMTC. The eMBB slice (RAN Slice ID=1) is used for general public mobile broadband services, one of the mMTC slices (RAN Slice ID=3) for small enterprise customers with IoT needs (e.g., transportation companies) and the other mMTC slice (RAN Slice ID=4) for a few customers with large scale deployments of IoT applications (e.g., utilities). In addition, MNO#A provides wholesale access to two 3rd party service providers, leasing capacity from its RAN infrastructure. One 3rd party provider is a PS communications operator

(MNO#B) that uses the contracted RAN slice (RAN Slice ID=2) for Mission Critical Push-To-Talk (MCPTT) and MCVideo services in a way that supplements the capacity of its own dedicated PS broadband network (PLMN#B). The second provider is an IoT communications provider (MNO#C – PLMN#C) specialized within the automotive sector, who exploits the leased RAN slice (RAN Slice ID=5) customized to serve the needs of connected car applications.

RAN deployed by a PS communications provider

A PS communications service provider (e.g., FirstNet in US [19]) deploys a dedicated RAN infrastructure (i.e., dedicated spectrum and RAN nodes) to serve the diverse needs of different PS agencies. In this respect, the RAN is configured in a first stage with two RAN slices, one for eMBB (RAN Slice ID=1) and another for mMTC (RAN Slice ID=2). The former is used for MCPTT and MCVideo services while the latter is used for asset tracking, automatic vehicle location, traffic signs and sensors, speed radars, etc.

In a second stage, sustained in the high reliability delivered by the PS RAN, the business of the PS service provider is extended towards the delivery of ultra-reliable MTC (uMTC) with extreme requirements on availability, latency and reliability to users in the areas of e.g., V2X communications for road safety (RAN Slice ID=3) and critical infrastructures monitoring applications (RAN Slice ID=4). In addition, in order to attain additional revenues, spare capacity in the PS RAN is leased to a commercial MNO (RAN Slice=5).

RAN deployed by a neutral host provider

The owner of a large venue (e.g., stadium, shopping mall, office building) deploys a neutral RAN infrastructure that is leased to MNOs in the form of RAN slices so that MNOs can provide service to their subscribers within the venue without the need to deploy their own infrastructure. The service is contracted by three nationwide MNOs so that three RAN slices (RAN Slice ID=1, 2, and 3) are configured, one per MNO. In addition, 2 RAN slices are established for local services, one slice (RAN Slice ID=4) to support security services (surveillance cameras, private communications) and another (RAN Slice ID=5) for remote monitoring and control of the installations (industrial applications). Finally, the venue owner deploys a local content delivery system to enrich the experience of its customers and to that end configures a new RAN slice (RAN Slice ID=6) optimized for Ultra High Definition video broadcasting.

The detailed illustrative view of the radio interface protocol architecture and the corresponding descriptors for the first scenario are provided, respectively in Table 1, in Figure 5 and Table 2. The overall configuration consists of five *RAN Slice Descriptors* that include a combination of five *L3 Slice Descriptors*, four *L2 Slice Descriptors* and three *L1 Slice Descriptors*. At the physical layer, 80% of the cell resources are allocated to *L1 slice #1* configured for eMBB services, 10% to *L1 slice #2* configured for mMTC services (which might be based e.g., on the use of legacy NB-IoT physical layer) and the remaining 10% allocated to *L1 slice #3* for the delivery of uMTC services (which might be based e.g., on a new 5G waveform optimized for vehicle-to-infrastructure communications, denoted as Car IoT in this example). The two RAN slices *(RAN Slice ID*=1 and 2) sitting on top of the L1 slice for eMBB access are configured to use the same MAC sub-layer instance (e.g., MAC-eMBB in Figure 5) and the same set of common logical channels (i.e., BCCH, PCCH and CCCH). On this basis, traffic isolation between *RAN Slice ID*=1 and 2 is enforced at both L2 and L3 levels by means of separate L2 slice descriptors and L3 slice descriptors. Note that, as described in Table 2, the *Slice Authorized Capacity* is given in terms of percentage of L1 slice consumed resources, aggregated GBR and maximum number of simultaneously active *UE Contexts*. A higher allocation priority and scheduling priority is given to *RAN Slice ID*=2 for PS

communications. Like the two RAN slices for eMBB, the two RAN slices for IoT services (*RAN Slice ID*=3 and *RAN Slice ID*=4) are also configured to use a shared MAC sub-layer instance (e.g., MAC-NB-IoT in Figure 5) and set of common logical channels. However, isolation is in this case only configured at L3 level (i.e., both RAN Slices uses the same L2 configuration descriptor). Finally, it can be noted that the Car IoT physical layer slice is configured to be only exploited by *RAN Slice ID*=5, so that there is no need of isolation at L2 and L3.

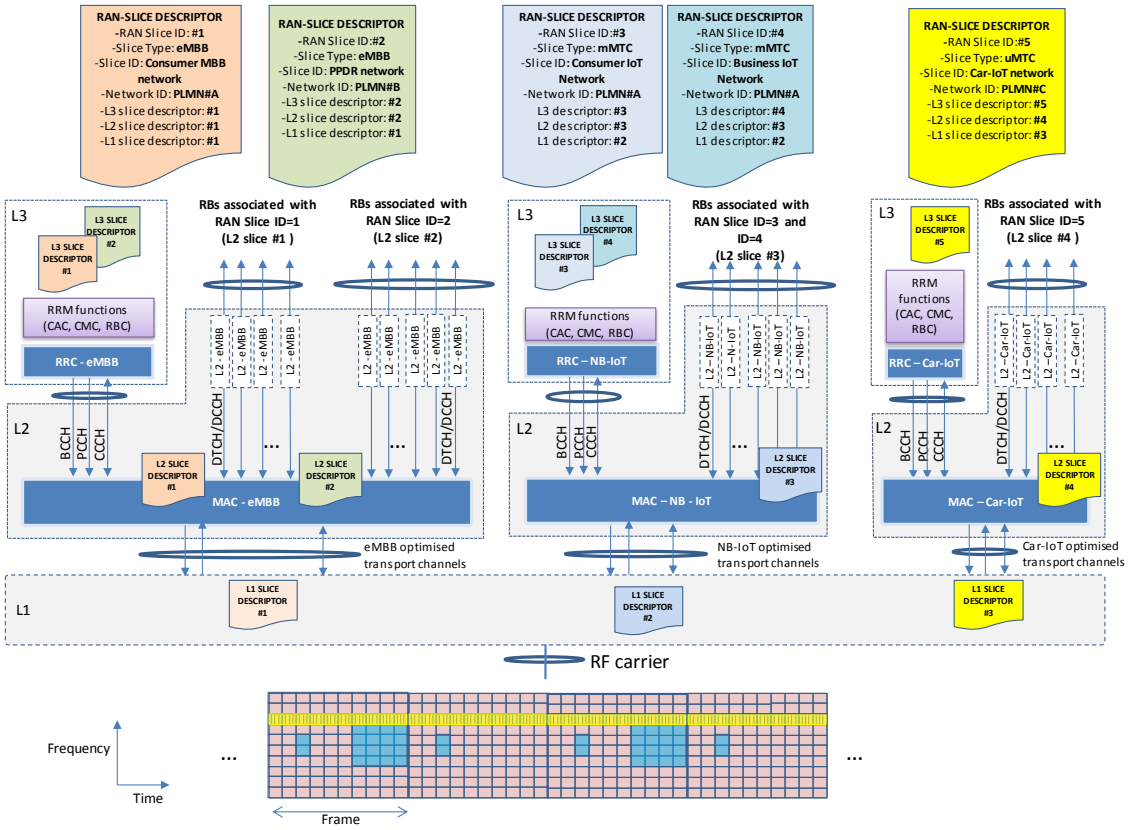

Figure 5: Radio protocol stack view of the RAN slicing configuration

Table 2: Illustration of the RAN slicing configuration descriptors

	ID #1	*ID #2*	*ID #3*	*ID #4*	*ID #5*
RAN Slice Descriptors	-*RAN Slice ID*=1 -*Slice Type*=eMBB -*Slice ID* = "Consumer MBB network" -*Network ID* = PLMN#A	-*RAN Slice ID*=2 - *Slice Type*=eMBB -*Slice ID* = "PPDR MBB network" -*Network ID* = PLMN#B	-*RAN Slice ID*=3 - *Slice Type*=mMTC -*SliceID* = "Consumer IoT network" -*Network ID* = PLMN#A	-*RAN Slice ID*=4 - *Slice Type*=mMTC -*SliceID* = "Enterprise IoT network" -*Network ID* = PLMN#A	-*RAN Slice ID*=5 - *Slice Type*=uMTC -*SliceID* = "Car IoT network" -*Network ID* = PLMN#C
L3 Slice Descriptors	*ID #1*	*ID #2*	*ID #3*	*ID #4*	*ID #5*
	-*Slice Authorized Capacity*[(1)]: [80%, 400 Mb/s, 200] -*Slice Allocation Priority*: 2	-*Slice Authorized Capacity*[(1)]: [20% , 100 Mb/s, 50] -*Slice Allocation priority*: 1	-*Slice Authorized Capacity*[(1)]: [50%, N/A, 2000] -*Slice Allocation Priority*: 1	-*Slice Authorized Capacity*[(1)]: [50%, N/A, 1000] -*Slice Allocation Priority*: 2	-*Slice Authorized Capacity*[(1)]: [100%, N/A, 500] -*Slice Allocation Priority*: N/A

	-*RRM policies*: Set by MNO#A	-*RRM policies*: Set by MNO#B			-*RRM policies*: Set by MNO#C
	-*RRC capability set*: eMBB profile -*Extended RRC features*: *Slice ID* broadcasting, Per-slice AB/LC, Per-slice cell (re-)selection		-*RRM policies*: Set by MNO#A -*RRC capability set:* NB-IoT profile [20] -*Extended RRC features*: *Slice ID* broadcasting, Per-slice AB/LC		-*RRC capability set:* Car-IoT profile -*Extended RRC features*: *Slice ID* broadcasting
L2 Slice Descriptors	*ID #1*	*ID #2*	*ID #3*		*ID #4*
	Slice-AMBR=1600 Mb/s -*Slice scheduling priority*=2 -*Slice Resource Utilization*=80%	-*Slice-AMBR*=400 Mb/s -*Slice scheduling priority*=1 -*Slice Resource Utilization*=20%	-*Slice-AMBR=N/A* -*Slice scheduling priority=N/A* -*Slice Resource Utilization*=100% -*L2 capability set*: Single HARQ process; Only RLC AM mode with simplified status reporting; No AS security (Non AS [NAS] security is used instead); PDCP operating in transparent mode.		-*Slice-AMBR=N/A* -*Slice scheduling priority=N/A* -*Slice Resource Utilization*=100% -*L2 capability set*: 6 HARQ processes; Only RLC AM mode with simplified status reporting; AS security.
	-*L2 capability set*: 6 HARQ processes; RLC TM, UM and AM modes with full status reporting; Access Stratum (AS) security.				
L1 Slice Descriptors	*ID #1*		*ID #2*		*ID #3*
	-*Radio resource allocation*: 80% of physical resources; transmission pattern with frame periodicity. -*L1 capability set:* Sub-frame duration: 1.0 ms Subcarrier spacing: 15 KHz Modulation: QPSK/16-64-256QAM		-*Radio resource allocation*: 10% of physical resources; transmission pattern with a periodicity longer than the frame duration. -*L1 capability set:* Sub-frame duration: 1.0 ms Single tone with 3.75 kHz spacing Modulation: QPSK		-*Radio resource allocation*: 10% of physical resources; transmission pattern with continuous allocation of a number of subcarriers. -*L1 capability set:* Sub-frame duration: 250 µs Subcarrier spacing: 30 KHz Modulation: QPSK
	-*L1 capability set:* OFDM-*based* waveform with scalable numerology. Windowing techniques to suppress inter-numerology interference [17]. Common frame structure of 10 ms.				

[(1)] *Slice Authorized capacity : [% of L1 slice resources, Aggregated GBR, maximum number of UE Contexts]*

2.3 Self-X Framework Northbound

The previous section has shown that the possibilities to configure the RAN slicing are vast. In a similar way, there are some further Self-X features in SESAME that require -or allow- inputs that are not directly provided by other SESAME components or Self-X features, but from external entities, like the VSCNOs. This section gives a short overview of which interactions between external entities and the Self-X framework are considered.

The SC is one of the elements that can be configured form the outside: The Self-X capabilities of the PNF (transmit parameter selection and automatic neighbour relations) are controlled by configuration parameters. These parameters may be set via the EMS GUI or via the northbound CM interface to the EMS. Currently, in the SESAME PoC, they are configured solely by the Small Cell Network Operator (SCNO) but it is envisaged that a commercial deployment would allow to provide for input from each VSCNO. Currently, for transmit parameter setting, the configuration parameters of the PNF specify aspects such as:

- How often and at what time a radio environment scan is performed;
- The frequency bands to scan;
- The desired coverage (and hence transmit power) of the PNF;
- Upper and lower transmit power bounds;
- A list of values from which the PNF may select its PRACH Root Sequence;
- A list of values from which the PNF may select its Physical Cell Identity (PCI).

These are discussed in more detail in Section 3.4.

In addition to the above, for Automatic Neighbour Relations (ANR), the configurable parameters of the PNF include:

- Whether to do a full scan of all configured frequency bands or constrain the scan to those on which neighbours were previously detected;
- A TX power threshold, above which detected neighbours are considered macro cells;
- A list of blacklisted carriers to which handout should not be attempted.

These are discussed in more detail in Section 3.5.

When multiple VSCNOs are introduced into the mix, the PNF may be required to consider inputs from each VSCNO such as:

- The desired coverage of the virtual cell. This will vary from one VSCNO to another, based on the area where they wish to provide service and the location of neighbour cells supporting their PLMN. In this case, the PNF will be responsible for balancing the requirements of the VSCNOs without compromising, *for example*, its target cell edge Signal to Interference and Noise Ratio (SINR).

- Specific, statically configured, neighbour cells provided by each VSCNO. The SESAME PoC has side-stepped this issue by only making use of neighbour cells detected by ANR. However, small cells typically have neighbour lists containing a combination of cells detected by a radio environment scan and those explicitly configured by the network operator (i.e., known to be present but not necessarily detected). In the case where multiple VSCNOs are stakeholders in the configuration of the PNF it is possible that each VSCNO may wish to specify a number of neighbours that are explicitly included in the

PNF's configuration. The EMS must provide an API to permit this and the PNF must contain the logic to resolve the demands placed on it by each virtual cell that it serves.

Apart from the configuration of the SC PNF, the 5G-EmPOWER also provides options on how to configure the RAN slicing. For the configuration of the RAN slices, the 5G-EmPOWER controller provides a set of RESTful APIs to enable creating, deleting and modifying a tenant/RAN slice in small-cell/eNB. This API is also the main resource for a tenant to modify the slice configuration. As any standard REST interface, this is based on different URL for GET/POST/PUT/DELETE methods, with the GET calls that do not require authentication. Specifically, the authentication method used in 5G-EmPOWER is based on basic HTTP authentication to limit the access to the RESTful resources. Two different account levels are possible: root and user. The standard RESTful API for 5G-EmPOWER has the following format

http://{username}:{password}@api-empower.create-net.org:8888/api/v1/,

where *api-empower.create-net.org:8888* has to be replaced with the hostname and port of the machine in which 5G-EmpOWER is locally installed. After the /api/v1/ different methods can be added to perform different actions. More information of 5G-EmPOWER API is made available in [22].

3 Technical details of the Self-X Functionalities in SESAME

In order to complete the description of SESAME's Self-X framework, this section provides in-depth technical details on the most relevant of the Self-X features introduced in Section 2.1, while at the same time explaining its mapping to the SESAME architecture and explaining its interactions with other Self-X features of the framework. Each of the description of a Self-X feature can be broken down into 4 parts:

- *A description of the feature*: Each Self-X feature is introduced, explaining its scope and what its duties are.
- *Details on the implementation*: Each Self-X feature can be implemented either as a VNF or a PNF and it has been designed for a specific architecture. This bullet point points out these details and also highlights its main characteristics.
- *Its mapping to the SESAME architecture*: Every Self-X feature has its place in the architecture, whether it is located in the Light DC, a CESC, the CESCM or even in another component of the SESAME architecture. Here the details on where the Self-X feature is located and whether it is a dSON, cSON or hSON feature can be found.
- *Its workflow within the SESAME architecture*: The workflow describes the interactions and the APIs used between Self-X features with other features and SESAME.

Some Self-X features also provide a performance assessment that evaluates how the features perform in specific test scenarios, extending the analysis that was presented in the previous deliverable D3.2.

3.1 Spatial Resource Scheduler for Self-Optimising Small Cell Networks

Already in [1] the concept of a spatial resource scheduler for a Self-Optimising small cell tier was introduced. The key idea was to use a two-level scheduler that can take advantage of the time-frequency diversity of the Physical Resource Blocks (PRBs), as well as take into account diversity in the spatial domain. Essentially, the two-level scheduler makes the attempt to enable a more efficient scheduling of the resources to alleviate the Inter-Cell Interference (ICI) problem, which is particularly crucial for cell edge users. To add also the concept of spatial reuse, the spatial scheduler relies on the well-known technique of Fractional Frequency Reuse (FFR) applied to small cell networks. Although small cells mainly operate in omni-directional mode, the idea is to enable cell sectoring also in this case. The possibility of applying fractional frequency reuse, particularly strict FFR technique, has shown already to be appealing to manage PRBs that can be allocated to the users [23]. Referring to the SESAME general architecture, which is made of CESCs, the principle consists of enabling an inner region in the CESC coverage area where frequency reuse one is used (FR_1), while the outer region can be partitioned in a variable number of sectors (or scheduling areas denoted as FR_n) depending on the beamwidth ϑ used. This approach is referred to as dynamic strict FFR. The higher the number of partitions (denoted by *N* for generality), the lower the bandwidth available in each one (see Figure 6).

Objective of the present study: Study the performance of centralized and distributed algorithms for a Self-Optimising network feature in the context of SESAME to exploit time, frequency and spatial diversity to schedule downlink packets. Under dynamic strict FFR, this is done determining for each user the best possible MCS, and the best way of partitioning the cell coverage area in order to relieve the ICI problem. The use of the term "dynamic" has to be

intended as the possibility to change the width of the inner region with frequency reuse one, as well the number of partitions in the outer region, in opposition to a static decision. Performance benchmark will be done with respect to CESCs that use frequency reuse one and an omni-directional antenna. In addition to the study previously presented, performance with and without power control is also addressed, as well as a heuristic approach is proposed and compared with the other techniques.

Summary of the approach: The key objective of adopting a two-level scheduler that operates on different time scale, i.e., the standard one millisecond MAC scheduler (or TTI scheduler), and the spatial scheduler that carries out resource scheduling over tens or hundreds of milliseconds (or even longer), is to enable the selection of the best possible MCS for each user, while alleviating the ICI problem and performing the usual scheduling of PRBs in the time-frequency domain. Two approaches were presented already in [1] taking the point of view of a reference link within a CESC. The first approach consists of a centralized spatial resource scheduler, which makes the attempt to optimize transmissions of downlink packets in the best possible way, relying on a global view of the whole CESC cluster state. The second consists of a distributed approach in which decisions are made for a scheduling area within the CESC.

Selected algorithms: The centralized spatial resource scheduler is modelled through to the combination of the Generalized Assignment Problem (GAP) and the Multi-Choice Knapsack Problem (MC-KP), two well-known NP-hard problems in combinatorial optimization theory ([23], [24], [25]). The combination of both yields the Multi-Choice GAP (MC-GAP) problem, which can be solved through computationally efficient algorithms. The distributed approach relies instead on solving the MC-KP in each cell sector (or alternatively a group of them).

The work presented in [1] has relied on generalizing previous research works [26] to model the probability of coverage for the inner frequency reuse one area. In specific, it was used the tool of stochastic geometry to model the coverage probability when aggregate interference affects the cell edge users, upon modelling the spatial distribution of CESCs over the two-dimensional plane according to a homogeneous Poisson Point Process (PPP) with intensity λ, under Rayleigh fading and different path-loss conditions (modelled through the path-loss exponent α). The other assumption is that each user will associate to the nearest base station. Under such general hypotheses, it was assumed the point of view of a reference downlink connecting a reference UE with the nearest serving CESC. Two thresholds were introduced, one to model the probability of coverage in the inner area (denoted by β_{FR}), and another one (denoted as β) to determine whether a downlink packet was successfully received by the reference UE. At start, the downlink communication is assumed affected by the ambient noise, as well as the aggregate interference that arise from other (uncoordinated) CESCs over the same frequency band of the reference link. Therefore, the SINR is simply written as follows:

$$\gamma \triangleq \frac{P_t g r^{-\alpha}}{P_\sigma + I} = \frac{P_{t0} g r r_0^{-\alpha}}{P_\sigma + \sum_n P_{tn} g r_n^{-\alpha}}, \quad (1)$$

where P_σ is the in-band power of the AWGN noise and I is the aggregate interfering process. On the other hand, g_n is the channel power fading coefficient for the nth interferer, P_{t0} and P_{tn} are respectively the transmit power of the reference link and the nth interferer, r_0 and r_n are respectively the reference link distance and the distance of the nth interferer from the reference UE.

As shown in [23] the coverage probability P_c can be defined upon setting a threshold β as follows

$$P_c \triangleq P_r\{\gamma > \beta | g_n, r_n\} \quad (2)$$

Under the hypothesis of using strict FFR for partitioning the reference CESC coverage, external and inner coverage probabilities can be respectively defined as shown below.

$$P_{e,FR} \triangleq P_r(\gamma > \beta | \gamma < \beta_{FR}) \quad (3)$$

$$P_{i,FR} \triangleq P_r(\gamma > \beta | \gamma > \beta_{FR}) \ .$$

Since the detailed analytical work was already provided in [23], only the most relevant results are reported herein below. It is anyway worth emphasizing that the probability of coverage is done to determine the boundary between the inner region with frequency reuse one and the outer region that is further partitioned.

Theorem 3-1: In an interference-limited network of CESCs where the spatial process is modelled through a PPP with intensity λ and strict FFR is used with N possible sub-bands, given a threshold β and values of the path-loss exponent α>2, the probability of coverage can be expressed as:

$$P_c = \frac{1}{1+N^{-1}\rho(\beta,\alpha)} \ , \quad (4)$$

with $\rho(\alpha,\beta) = \beta^{-2/\alpha}\frac{2\pi\csc(2\pi/\alpha)}{\alpha} - 2F_1\left(1,\frac{2}{\alpha},\frac{(2+\alpha)}{\alpha},-\beta^{-1}\right)$, where the function $2F_1$ stands for the Hypergeometric function.

Upon following the same general assumptions stated in Theorem 3-1, the probability of coverage for users camping in the external region when FFR is applied carries the following expression

$$P_{FFR} = \quad (5)$$
$$\frac{P_c(\beta,\alpha,\lambda,N)}{1-P_c(\beta_{FR},\alpha,\lambda,1)} - \frac{\pi\lambda\int_0^\infty \exp\left(-\pi\lambda y\left(1+2\xi(\beta,\beta_{FR},\alpha,N)\right)\right)\times\exp\left(-\frac{(\beta+\beta_{FR})P_\sigma}{P_t}y^{2/\alpha}\right)dy}{1-P_c(\beta_{FR},\alpha,\lambda,1)} \ .$$

In an interference-limited network in which $P_\sigma \cong 0$, the above expression can be further simplified after some mathematics in the following way for $\beta \to \beta_{FR}$

$$P_{FFR} = \frac{P_c(\beta,\alpha,\lambda,N)}{1-P_c(\beta_{FR},\alpha,\lambda,1)} - \frac{1/\left(1+2\xi(\beta,\beta_{FR},\alpha,N)\right)}{1-P_c(\beta_{FR},\alpha,\lambda,1)} \ , \quad (6)$$

where

$$\xi(\beta,\beta_{FR},\alpha,N) = \frac{\beta_{FR}}{2\alpha(1+\beta_{FR})}\left[-\frac{\beta_{FR}\left(2-2\alpha+(\alpha-2)(1+\beta_{FR})\right)}{\alpha-1}\times\right. \quad (7)$$
$$2F_1\left(1,2-\frac{2}{\alpha},3-\frac{2}{\alpha},-\beta_{FR}\right) + 2\left(1+\frac{1}{N}\right)\times$$
$$\left.\frac{\left(-2+\alpha+2(1+\beta_{FR})2F_1\left(1,\frac{\alpha-2}{\alpha},2-2\alpha,-\beta_{FR}\right)\right)}{\alpha-2}\right] \ .$$

Under strict FFR assumption, the number of resource blocks (out of a total N_b) that are made available to the users camping in the outer region can be expressed as follows:

$$N_{ext} = \left(1 - P_{FFR}(\beta_{FR})\right) \times N_b \ . \quad (8)$$

Theorem 3-2: For the sake of computing the probability of a packet success (P_s), the working assumption is to consider Rayleigh fading for any value of the path-loss exponent α>2 affecting

the reference link of size r_0 in the presence of AWGN noise with in-band noise power N_0W and for a detection threshold β. In case aggregate interference is completely suppressed through dynamic strict FFR (such as in the centralized two-level scheduler), P_s can be written as:

$$P_s = \exp\left(-\frac{\beta}{P_{t0}} N_0 W r_0^{\alpha}\right), \quad (9)$$

where P_{t0} is the transmitted power used in the reference link, and the above expression comes from the fact that a Rayleigh distributed amplitude fading implies an exponentially distributed power fading.

In case residual interference is still present due to an incomplete suppression through strict FFR (such as in the decentralized two-scheduler approach), the expression of the probability of success can be generally derived to include also power control. Relying on the work done in [27], it is possible to introduce a parameter p, which is used to enable the power control so that $P_t=pP_{t0}$, and a parameter $P=gP_t$, which takes into account both fading and transmitted power of the interfering CESCs. The maximum transmitted power common to all CESCs is instead denoted as P_{t0}. It is interesting to observe that the power control parameter can also be a way of thinning the original PPP by scaling the intensity to λp. Therefore, the parameter p is included in the intensity of the equivalent thinned spatial process. In this study, the power control is intended a way of compensating the path-loss between an interferer and the UE in the reference downlink transmission. Therefore, the transmitted power is modelled as a random variable $P_t=r^{\alpha}$. It is easy to prove that such transmitted power follows a Weibull distribution

$$P_r\{P_t \leq x\} = 1 - \exp(-\lambda\pi x^{\nu}), \quad (10)$$

where ν depends on the path-loss exponent α. It is hence straightforward to show that

$$E(P_t^{\nu}) = \frac{\Gamma\left(1+\frac{\nu}{2/\alpha}\right)}{(\pi\lambda)^{\frac{\nu}{2/\alpha}}} = \frac{1}{\pi\lambda}. \quad (11)$$

where it was used $\nu = \alpha/2$. Separating fading from the transmitted power, it is possible to obtain

$$P_s = exp\left(-\lambda\pi p r_0^2 E\left(P^{\frac{2}{\alpha}}\right)\Gamma\left(1-\frac{2}{\alpha}\right)\beta^{\frac{2}{\alpha}}\right) = \quad (12)$$

$$= exp\left(-\lambda\pi p r_0^2 E(P_t)E(g^{2/\alpha})\Gamma\left(1-\frac{2}{\alpha}\right)\beta^{\frac{2}{\alpha}}\right) =$$

$$= exp\left(-p r_0^2 \Gamma\left(1+\frac{2}{\alpha}\right)\Gamma\left(1-\frac{2}{\alpha}\right)\beta^{\frac{2}{\alpha}}\right).$$

It is interesting to notice that with p=1 (no power control) the probability of success reduces to

$$P_s = \exp\left(-\lambda\pi r_0^2 \frac{2\pi/\alpha}{\sin(2\pi/\alpha)}\beta^{2/\alpha}\right), \quad (13)$$

in which the transmitted powers of all CESCs is just set to the maximum, and thus they do simplify each other. In the equation (1) it has also been used the Euler's reflection formula, if compared to the last in equation (12).

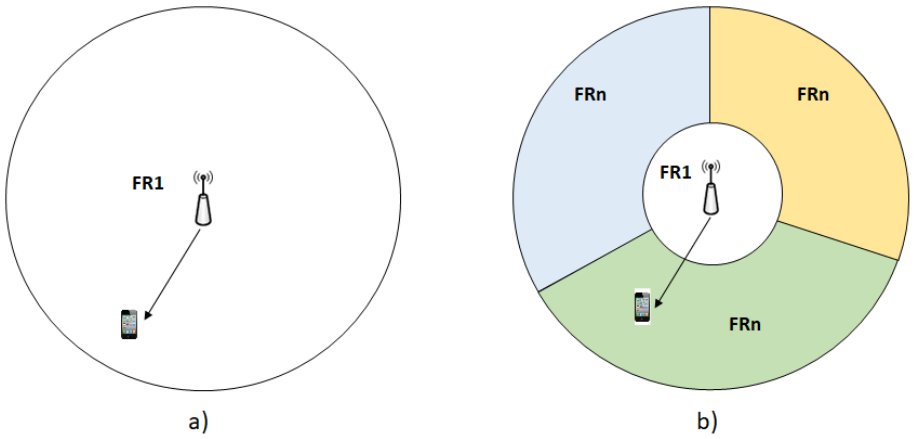

Figure 6: Reference scenario for applying dynamic strict FFR

Proposed heuristic solution to the MC-GAP problem

Since the MC-GAP problem is known to be NP-hard, it is relevant to provide a heuristic solution to solve it that can reduce the computational complexity, while providing a satisfactory solution to the multi-choice assignment problem. To explain the heuristic, it is worth reminding that a scheduling area which is obtained after partitioning the CESC coverage is also referred to as *bin*. A packet to be transmitted is also referred to as *item*, interchangeably. On the other hand, a *configuration* simply refers to a feasible MCS to use for a downlink packet. The heuristic algorithm is explained through an example to provide an intuitive understanding of the procedure. All values that are provided do not for the moment reflect the real values associated to a specific MCS or scheduling area but have just been selected for the sake of the explanation.

As shown in Figure 7, the example procedure to illustrate the heuristic includes three scheduling areas (or bins) each having a capacity of B_1 = 15 PRBs, B_2 = 10 PRBs and B_3 = 9 PRBs, *respectively*. For simplicity, *k*=1,2,3,4 packets are to be transmitted, and *m*=1,2,3 different configurations for sending packets in a scheduling area are possible. As in the MC-GAP problem, the *i*th packet is transmitted in bin *j* with the required number of PRBs (weight w_{ij}^m) and an associated profit (p_{ij}^m) when configuration *m* is used. A more formal definition of profit and weight will be provided in Section 3.4.3. The tables in Figure 7 show the weights and profits of the four packets in each bin and configuration.

- The first step consists of finding the configuration that gives the best profit in each bin for an item. For the selected configuration, the weight of each item is highlighted in the weight table with the light silver colour. Referring to the table of weights and profits in Figure 7, the thick black boxes show the item configuration (the pair weight, profit) that yields the best profit in different bins and the corresponding weight. To be more precise, in the first bins items 3 and 4 maximize their profit with an associated weight.
- The second step consists of checking whether the bin's capacity is violated or not, as in the MC-GAP problem formulation. To do that, the weights of all *k* packets for the selected configuration in a bin are added and compared against the bin size. To be concrete, the sum of weights of items 3 and 4 in bin 1 is 4+12 = 16 > 15, which violates the bin size capacity. For bins 2 and 3 it happens instead that B_2>3 and B_3>3, respectively for item 2 and 3. The corresponding sum profit is 25+36+30+32 = 123. Anyway, the bin 1 capacity constraint is violated, and an alternative way of packing must be found, though less efficient. Clearly, the best case would have been if all bin capacities could have been fulfilled, thus yielding the optimal packing for the given problem. Since the search of the

problem solution would have stopped, this case would have been also the one with lower computational complexity since no further operations would have been required.

- The third sept is to satisfy the capacity constraint of those bins that violate it.
    - The procedure consists of checking the second-best profit with a lower weight within the bin for either item 3 or 4. It is changed first the configuration of item 3 to configuration 2 in bin 1 (shown in Figure 7 in blue right slanted pattern) with item size 3 and profit 15, while keeping the configuration of item 4 unchanged. Then, this configuration is compared with the best way of packing item 3 with the best profit in the other bins (the best configuration in bin 2 gives weight 3 and profit 18 as shown with the black dash edge rectangle; for the same item bin 3 shows weight 2 and profit 4). Upon doing the decision of packing item 3 in bin 2, all the weight constraints are satisfied: B_1>12 for item 4, B_2>3+3 for items 2 and 3 and B_3>3 for item 1. The overall profit of packing all the items in all the bins becomes 36+30+18+32 = 116.
    - The procedure is iterated to check if the sum profit can be further improved by only changing the configuration of item 4, while keeping item 3 in bin 1 with configuration 3. For item 4 the configuration is changed from configuration 1 to 2 (shown in blue left slanted pattern with weight 8 and profit 32). Then the profit of item 4 in bin 1 is compared with the other bins, and it can be found out that configuration 2 in bin 1 yields the best profit. At a second stage, it is checked whether the bin 1 capacity constraint is still satisfied with the new configuration for item 4 (i.e., B_1>4+8 = 12). The new profit using this configuration becomes 25+32+30+32 = 119, which shows to be the optimal feasible solution to the packing problem: [I_1, C_1, B_3], [I_2, C_2, B_2], [I_3, C_3, B_1] and [I_4, C_2, B_1].

It is worth noticing that the heuristic might not always converge to the optimal solution quickly. In such cases, the idea is to converge towards a sub-optimal solution breaking the search after a certain number of feasible configurations was found, and selecting the best. Generally, the heuristic takes the advantage of relaxing in the first place the hard constraint on the bin capacity used in the MC-GAP since it first searches the best profit of every item in each bin, and only afterward it checks the combinations that allow to fulfil the bins capacity constraint. This relaxation allows in some cases to achieve even better profit than the standard MC-GAP problem.

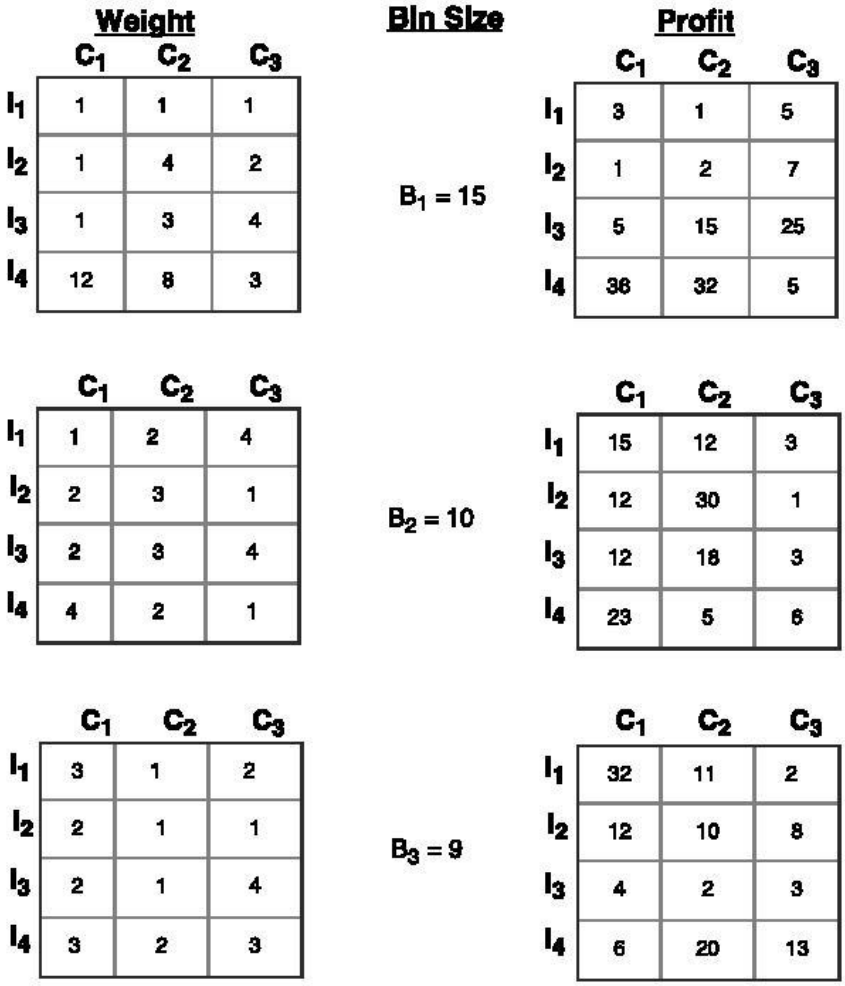

Figure 7: Weight-profit tables for four items, three bins and three configurations

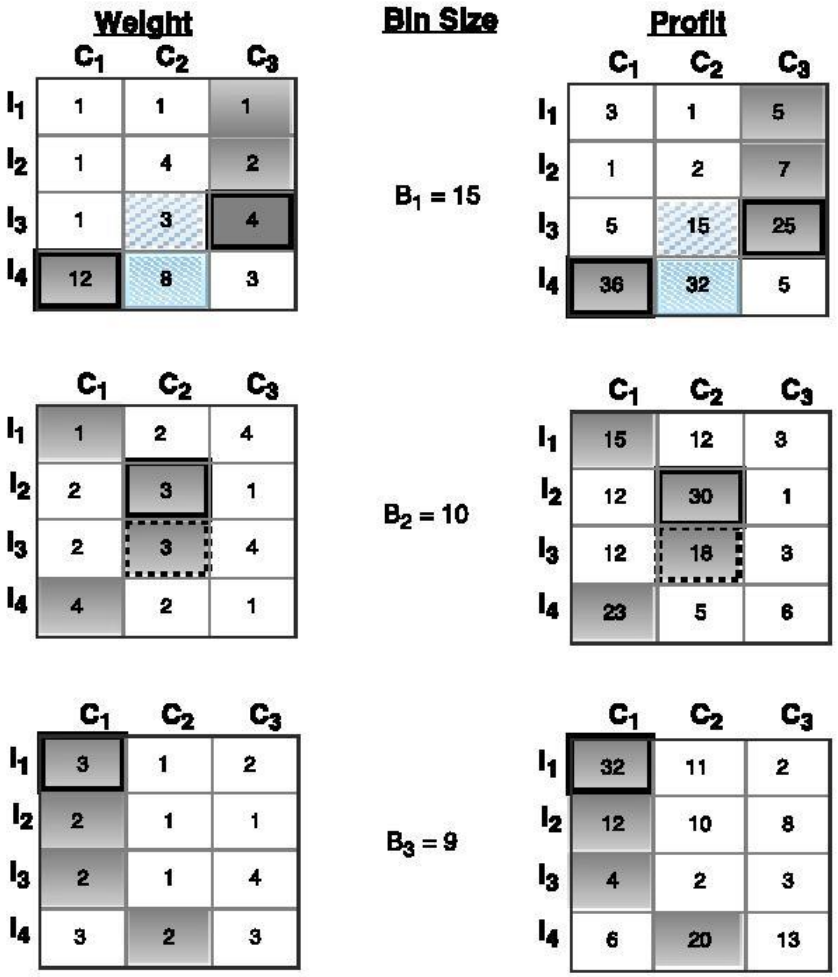

Figure 8: Weight and profit tables highlighting the different steps of the heuristic

3.1.1 Implementation details

The implementation of this type of Self-Optimising network feature changes depending on whether the algorithm implementing the spatial scheduler runs in the EMS or in a CESC, hence implying whether the decision on how to partition the CESC coverage area and the search of the best possible MCS for the downlink packets are made in centralised (cSON) or distributed (dSON) way. A centralised decision can be made also at the level of the Light DC, for instance for a group of CESCs but not restricted only to this case. Indeed, the Self-Optimising function can be also implemented as a VNF depending on whether the algorithm runs in the Light DC (virtualised environment) or not. More specifically, the algorithm could run within a single micro-server of the Light DC for a single CESC or take full advantage of the edge-cloud environment and make decisions for a group of CESCs. The reconfiguration of the radio parameters after the algorithm has determined the optimal solution does not require any firmware update at the PNF side. In case the Self-Optimising function is located in the EMS, the standard management interface

TR069[3] and the related data model TR196[4] can be used for the purpose of acquiring statistics from the PNF. On the other hand, in case the Self-Optimising function is executed in the CESC or in the Light DC, it is foreseen that an additional module for collecting statistics from the CESCs should be deployed to provide inputs to the spatial scheduler. In this latter case, a higher degree of flexibility in upgrading, migrating and deploying a new upgraded function is foreseen. On the other hand, when the function is placed in the EMS software upgrades can be done less frequently, and hence some flexibility is lost.

Given the complexity of the proposed two-level scheduler, the results that will be shown are mainly based on Matlab simulations to provide evidence of the performance that can be achieved.

3.1.2 Performance Assessment

For the sake of showing the results, it is worth reminding the metrics that are used in the MC-GAP and MC-KP problems

Definition 3-1: The weight and profit of the *i*th packet scheduled for transmission in bin *j* and MCS *m* can be written as follows

$$w_{ij}^{k} \triangleq \frac{L_p}{2 \times 7 \times 12 \times b_m \times r_{cm}} \qquad (14)$$

$$P_{ij}^{k} \triangleq \frac{b_m \times r_{cm} \times (1 - P_e) \times d_f}{b^{[64QAM,3/4]} \times r_c^{[64QAM,3/4]}}$$

where L_p is the packet measured in bits, b_m and r_{cm} denote respectively the number of bits and code rate used in the *m*th MCS, d_f denotes the useful fraction of a packet after removing the overhead and P_e is the packet error probability. The terms $b^{[64QAM,3/4]}$ and $r_c^{[64QAM,3/4]}$ denote the maximum allowed modulation and coding scheme (MCS_{max}).

Definition 3-2: The probability of packet blocking is the ratio between the number of dropped (N_d) to the number of packets awaiting transmission (N_t), where N_d is defined as the difference between N_t and the number of packets successfully transmitted N_s. Hence the blocking is as follows

$$P_B \triangleq \frac{N_d}{N_t} = \frac{N_t - N_s}{N_t}. \qquad (15)$$

N_d is mainly affected by the proportion of radio resources allocated in each bin, i.e., it directly depends on the FFR strategy. For the sake of showing results subsequently, the following metrics will be studied.

---

[3] TR-069 (Technical Report 069) is a technical specification that defines an application layer protocol for remote management of end-user devices. It was published by the Broadband Forum and entitled CPE WAN Management Protocol (CWMP).
For more details see, *inter-alia*: *https://en.wikipedia.org/wiki/TR-069*

[4] TR-196 (Technical Report 196) is a Broadband Forum technical specification. Its official title is "Femto Access Point Service Data Model". The purpose of this Technical Report is to specify the Data Model for the Femto Access Point (FAP) for remote management purposes using the TR-069 CWMP. For more details see, *inter-alia*: *https://en.wikipedia.org/wiki/TR-196*

Definition 3-3: The normalized spectral efficiency (η) of the system that applies dynamic strict FFR is the number of useful bits per transmitted packet with respect to the best possible MCS (i.e., MCS_{max}). It also depends on the packet overhead (d_f) and loss of packets (packet error probability P_e) due to the radio channel propagation as shown below

$$\eta \triangleq \frac{b \times r_c \times (1-P_e) \times d_f}{b^{[64QAM,3/4]} \times r_c^{[64QAM,3/4]}}. \quad (16)$$

Definition 3-4: The aggregate throughput (S), the summation of all achievable profits for the packets to transmit is defined as follows

$$S \triangleq b^{[64QAM,3/4]} \times r_c^{[64QAM,3/4]} \times \frac{2 \times 7 \times 12}{TTI} \sum_{k=1}^{K} \eta_k \times N_{bk}, \quad (17)$$

where the first two terms on the right-hand side of the equal sign are de-normalizing factors, and η_k is the achievable spectral efficiency for the MCS assigned to the kth packet (i.e., the profit). The term N_{bk} denotes instead the number of resource blocks (i.e., the weight of a packet) that are necessary to transmit the kth packet. The term $2 \times 7 \times 12$ is the LTE specific parameter in one TTI (i.e., one millisecond).

Definition 3-5: The optimal beamwidth θ^* used to partition the outer cell coverage is the beamwidth that maximizes the aggregate throughput (or the normalised spectral efficiency) of the network. In other words, the beamwidth that maximises the profit in the MC-GAP or MC-KP problems.

Definition 3-6: The optimal threshold β^* to partition the CESC coverage in inner and outer areas is the SINR threshold that maximises network performance. In particular, the optimal threshold is computed through a weighted mean as follows

$$\beta^* \triangleq \frac{\sum_{i=1}^{\kappa} \beta_i \omega_i}{\sum_{i=1}^{\kappa} \omega_i}, \quad (18)$$

where β_i denotes the optimal SINR threshold that is observed with frequency ω_i out of a total number of simulations. In addition, κ denotes the number of groups of threshold values that recur during the simulations.

Table 3: Simulation parameters

Parameter	Value
System bandwidth (W)	20 MHz
N_b	100
Transmit power (P_t)	24 dBm
SINR threshold (β) for α=2.1	[-10 - 5]
SINR threshold (β) for α=4	[-5 - 10]
Beamwidth (ϑ)	$\{90^\circ, 120^\circ\}$
MCS (modulation, code rate)	[QPSK, 1/2], [QPSK, 3/4], [16QAM, 1/2], [64QAM, 1/2], [64QAM, 2/3], [64QAM, 3/4]
SINR range [dB]	[0 - 6.5]; [6.5 - 7.2]; [7.2 - 13.6]; [13.6 - 18]; [18 - 21]; [21 - 25]
d_f	0.75

Packet size (L_p)	{32, 128, 256} bytes
PPP intensity (λ)	0.15
Number of simulations	50

For the sake of showing results, the general setup is as in [23] with the simulation parameters that are shown for convenience in Table 3. Dynamic strict FFR technique is adopted to partition the CESC coverage area and correspondingly the system bandwidth. The MC-GAP problem is used to model the centralised scheduling of packets, whereas the MC-KP is used to model packet scheduling in each scheduling area. The goal of the spatial scheduler is that to determine the best way of dividing the CESC coverage area along with the selection of the best MCS for the packets to transmit. Results are shown for a reference CESC and compared also with the case of omni-directional antenna (i.e., when FFR is not in place) with and without power control when the optimisation performed by the spatial scheduler is done whereby the MC-KP algorithm. The weight and profit that have been mentioned earlier in the section are computed as shown in equation (14).

The probability of a packet success was already derived in equations (9), (12) and (13) respectively for the cases without interference (i.e., ICI problem completely relieved using the centralised approach), with interference and power control (i.e., the ICI problem cannot be completely relieved while using the decentralised algorithm), and with interference without power control. Furthermore, the results of the heuristic solution are also provided in the remainder.

While the spatial scheduler makes the attempt to transmit packets with the best possible MCS for a given spatial partitioning of the reference CESC coverage, the TTI scheduler selected in the simulations is the simple round-robin mechanism. It was already discussed in [23] that other choices are possible, such as the proportionally fair scheduler or the best CQI scheduler, to make a few examples. It is worth stressing that simulations are carried out selecting for each packet the associated SINR by drawing a randomly generated value within the wide SINR range [0, 25] dB, and selecting in this way the MCS to use in the transmission depending upon which interval it falls within. The overhead factor d_f was assumed equal to 0.75, whereas the spatial density of the PPP was set to 0.15. Simulations results are shown against the increasing number of generated packets, and for each number of packets 50 simulations were run for the sake of collecting statistics.

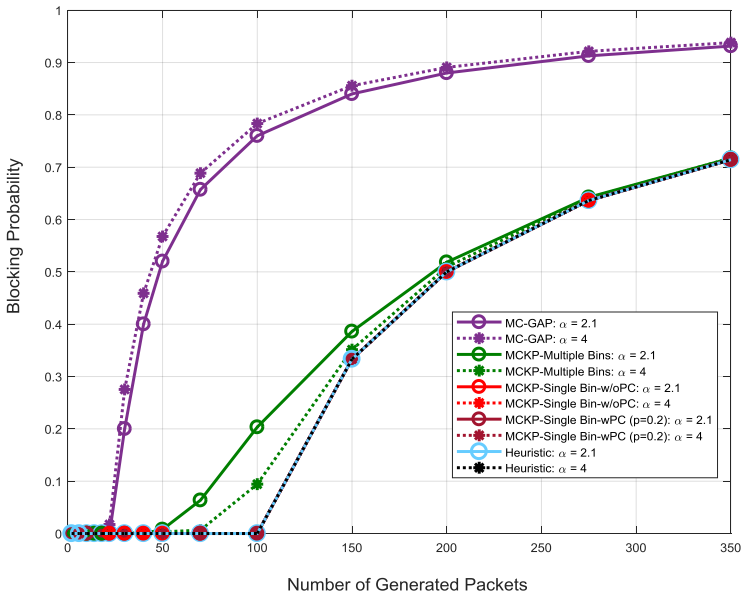

Figure 9: Average blocking probability versus the number of packets to serve for α={2.1, 4} and L_p=32 bytes. Results are provided for the MC-GAP, the MC-KP in single and multiple bins with and without power control, as well as with the heuristic.

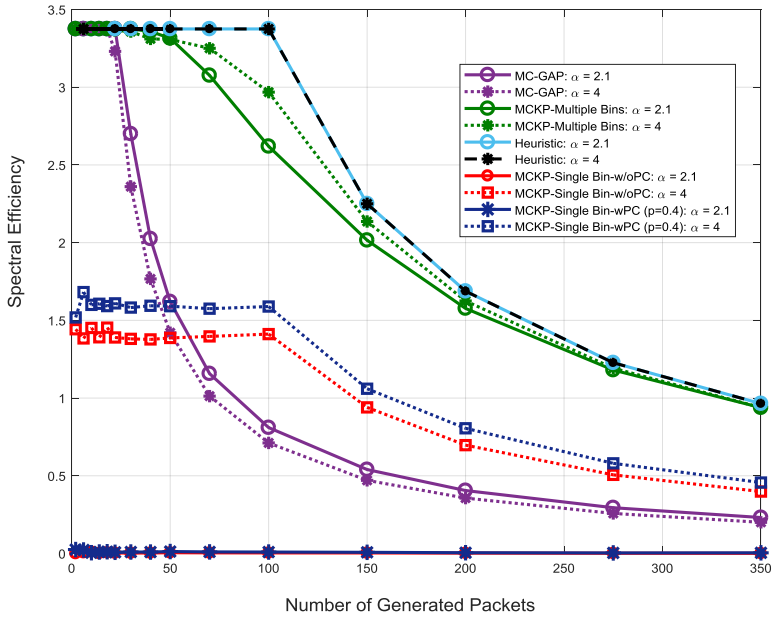

Figure 10: Average spectral efficiency versus the number of generated packets for α={2.1, 4} and L_p=32 bytes. Results are provided for the MC-GAP, the MC-KP in single and multiple bins with and without power control, as well as for the heuristic.

Figure 9 shows the average packet blocking probability (provided in equation (15)) and Figure 10 the average spectral efficiency (as in equation (16)) computed for a packet size of 32 bytes and values of the path-loss exponent α equal to 2.1 and 4. The results compare the standard MC-GAP and the heuristic algorithm that approximates the solution of the MC-GAP problem.

Furthermore, the MC-KP is applied to the cases of single and multiple bins. It is worth reminding that the term bin is alternatively used to denote a scheduling area. In the single bin case (i.e., when no FFR is used), power control is applied as a simple method to relieve ICI. The figure shows that the MC-GAP exhibits the highest blocking probability. This is a limit of the MC-GAP problem since the solution makes the attempt to schedule as many packets as possible even within in the scheduling area with the smallest number of PRBs when strict FFR is applied. The other packets are consequently dropped. All other approaches show lower blocking and when comparing the average spectral efficiency that can be achieved through the different methods, the heuristic solution exhibits the highest for any value of the path-loss exponent α. The MC-GAP and the MC-KP also show high spectral efficiency, whereas power control with p=0.4 performs almost as poor as the no power control case. Clearly, smaller values of p could be used but it is out of scope of the current study the optimization of the power control parameter. The results shown in Figure 9 and Figure 10 allow to conclude that if packets are transmitted without the spatial scheduler, they are transmitted losing the possibility to use higher order modulations. Hence, the spatial scheduler has the merit of optimising the cell partitioning, as well as the MCS to schedule packets, thus yielding higher performance. The heuristic can also improve the blocking probability since it relaxes the constraint of the MC-GAP to schedule as many packets as possible even within the scheduling area with the smallest capacity. It is also worth noticing that even without the spatial scheduler, disciplines such as the best CQI scheduler can improve the performance with respect to round robin. Anyway, the possibility to rely on a dedicated Self-Optimising feature is envisaged to perform better.

Figure 11 and Figure 12 show the average throughput (shown in equation (17)) that can be achieved while varying the number of generated packets (i.e., CESC load). Figure 11 shows the throughput of the MC-GAP, the heuristic and the MC-KP applied to multiple scheduling areas for a path-loss exponent α=2.1 and different packet size (32, 128 and 256 bytes). Figure 12 compares the different techniques for an increasing load and for α equal to 2.1 and 4 with a packet size of 32 bytes. The results shown in these two figures are consistent with the evidence found in Figure 9 and Figure 10. The heuristic solution exhibits the highest performance, whereas the MC-GAP is penalised by a packet drop higher than the other techniques. Figure 12 shows also the case with power control and p=0.2, showing quite clearly that an optimised power control scheme is already beneficial to improve the performance, and might be even a mandatory solution if other more sophisticated techniques are not implemented. Similar to the results of the spectral efficiency, when the interference cannot be completely relieved, a path-loss exponent higher than free space (quite common in indoors) is beneficial to shield from the ICI problem, thus yielding a higher throughput.

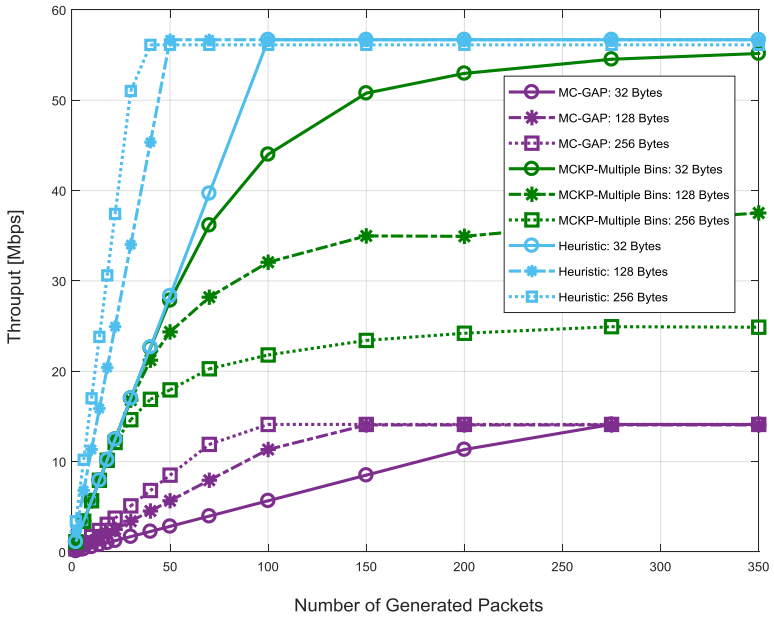

Figure 11: Average throughput versus the number of generated packets for α=2.1 and L_p={32, 128, 256} bytes. Results are provided for the MC-GAP, the MC-KP in multiple bins and for the heuristic.

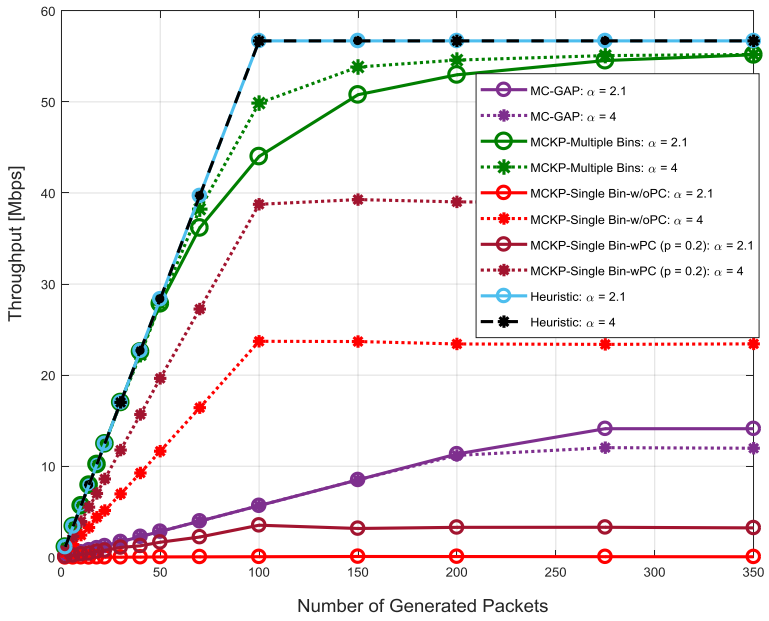

Figure 12: Average throughput versus the number of generated packets for α={2.1, 4} and L_p=32 bytes. Results are provided for the MC-GAP, the MC-KP in single and multiple bins with and without power control, as well as for the heuristic.

Figure 13 shows the optimal threshold θ^* to partition the reference CESC coverage area in inner and outer regions for values of the path-loss exponent for α=2.1 and 4 and a packet size 32 bytes. Results are shown against the number of generated packets for the standard MC-GAP

problem, the heuristic solution that approximates the MC-GAP and the MC-KP applied to multiple scheduling areas. The figure highlights that for all the three techniques and for each selected value of the path-loss exponent, convergence to an optimal threshold is reached. Furthermore, the figure shows that the MC-GAP exhibits the highest threshold for α=4 and the lowest for 2.1. It is also worth noticing that the MC-GAP exhibits the largest spread between the two values. Indeed, the MC-GAP makes the attempt to relieve completely the aggregate interference that arises from the surrounding downlink transmissions, which largely affects the search of a global optimised solution. This is reflected clearly in the low value of θ^* for α=2.1, and the high value for α=4. In the latter case, ICI is already reduced by a more severe propagation environment that attenuates not only the useful signal but interference as well. The heuristic solution behaves similar to the MC-GAP although with a smaller spread. Finally, the MC-KP shows to be least affected by the propagation environment since it does not target the search of a globally optimised solution. Comparing the three methods for α=2.1, the heuristic shows the highest threshold while the MC-GAP the smallest. This implies that applying the heuristic method, the inner coverage turns to be smaller than in the MC-GAP. On the other hand, for α=4 the MC-GAP exhibits the highest threshold, whereas the solution to the MC-KP problem the smallest. Hence, in this case, the inner coverage area for the MC-GAP is smaller than that in the MC-KP and of the heuristic solution.

Figure 14 and Figure 15 show the histogram of the optimum beamwidth ϑ^* choosing between the two candidate values of 90° and 120° for the MC-GAP and MC-KP, respectively. The results are shown for a path-loss exponent α=4 and packet size L_p=32 bytes, while increasing the number of generated packets. Figure 14 shows that when solving the MC-GAP in correspondence of a low number of packets that wait to be transmitted in downlink, a cell partitioning with a beamwidth of 90° (four sectors to partition the outer region of the reference CESC coverage) provides the best choice. On the other hand, as soon as the number of packets to transmit is increased, the best choice steadily converges towards a beamwidth of 120° (three sectors to partition the CESC coverage). The situation changes when solving MC-KP problem in the decentralised approach as shown in Figure 15. As shown by this figure, as the number of packets to transmit is increased not always the beamwidth of 120° is the best choice but also the 90° value is automatically selected to optimize the performance, although the choice of partitioning the CESC coverage with three sectors is still dominant. This difference is mainly due to the fact that the MC-GAP offers a way of modelling the optimization problem leveraging on the global view of the CESCs and the related neighbouring interference. Vice versa, the MC-KP is used to carry out local decisions that do not take advantage of the global information. Anyway, despite the MC-KP is in principle sub-optimal with respect to solving the MC-GAP problem, it does provide significant advantages in terms of throughput and spectral efficiency since it does not suffer the limitation inherent to the MC-GAP problem formulation, which causes a high packet blocking probability.

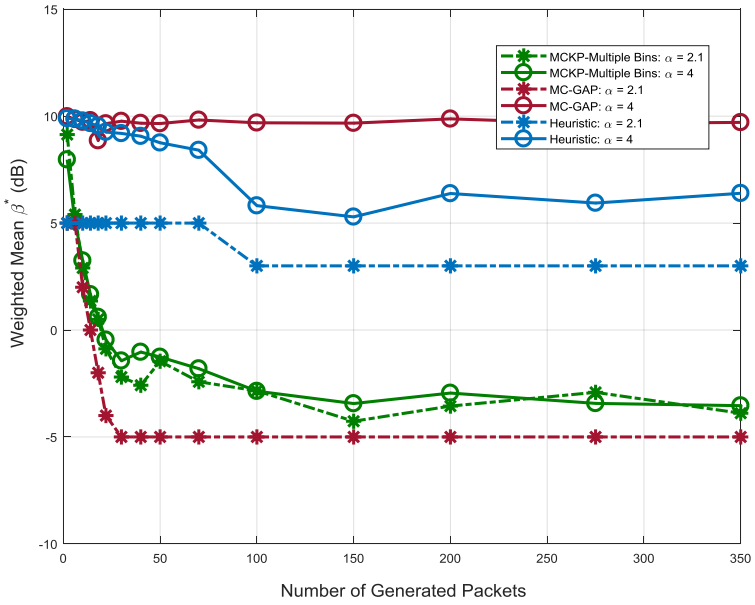

Figure 13: Optimal threshold versus the number of generated packets for α={2.1, 4} and L_p=32 bytes for the MC-GAP, MC-KP with multiple bins and the heuristic solution.

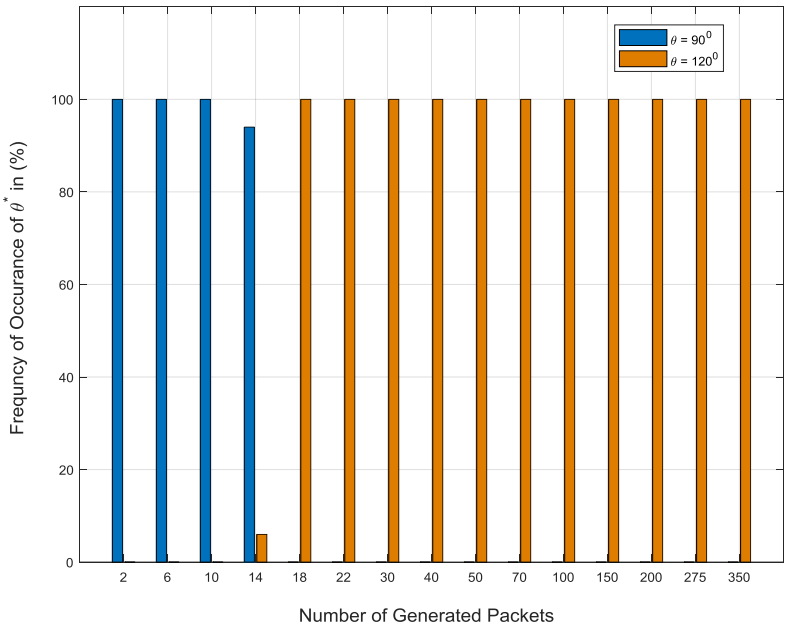

Figure 14: Histogram of the frequency of the selected beamwidth values versus the number of generated packets for α=4 and L_p=32 bytes when the MC-GAP algorithm is solved.

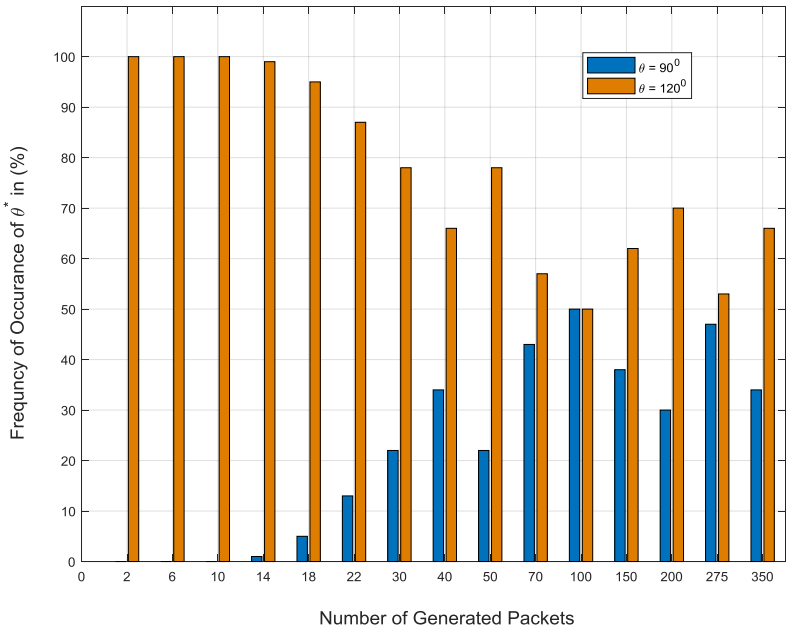

Figure 15: Histogram of the frequency of the selected beamwidth values versus the number of generated packets for α=4 and L_p=32 bytes when the MC-KP algorithm solved.

3.1.3 Mapping to the SESAME Components

The mapping of the two-level scheduler in the SESAME system architecture is shown in Figure 16. Compliant with the architecture, the CESCM includes all the necessary modules to perform management operations, as well as the level of orchestration typical of the ETSI MANO architecture [2]. The Light DC provides the virtualization environment (i.e., NFVI in ETSI terminology) where to run virtualized functions at the network edge, whereas the PNFs are connected to the micro-servers in a one-to-one fashion to form the CESC cluster. Different EMS that perform FCAPS operations are present, including the EMS PNF that relies on the management interface TR069, a standard way to collect statistics from base stations and small cells. Performance monitoring is a crucial part to enable Self-Optimization of the radio resources. Statistics such as cell or slice load, MCS and number of users within a cell or a slice are collected and averaged over a period of time to produce a time average. Further studies on such a time period of observation is out of scope of the current work.

Figure 16 shows indeed that dedicated functions can be deployed in the CESCs or in the Light DC for the sake of monitoring the performance of individual CESCs, a group of them or the whole cluster. This reflects the fact that the SESAME system is flexible and well designed to accommodate both centralized and distributed resource scheduling strategies. Particularly, the two-level scheduler is made of the spatial scheduler and the MAC resource scheduler. These two levels of scheduling operate on a different time scale with the MAC performing on a TTI basis (i.e., 1 ms). The spatial resource scheduler shall always operate on a longer time scale (e.g., hundreds of millisecond) relying on performance metrics that can be aggregated for a CESC, on a per tenant basis (i.e., VSCNO) or for the whole CESC cluster (i.e., infrastructure provider). The per tenant monitoring may involve one or more CESCs and reside in the Light DC as a VNF to enable both distributed and centralized spatial scheduling, whereas the overall infrastructure monitoring is located at the EMS level to enable centralized resource management and it cannot

be deployed as a VNF. Furthermore, depending on whether different functional splits are used (other the S1 virtualization) also the TTI scheduler can optionally reside in the CESC.

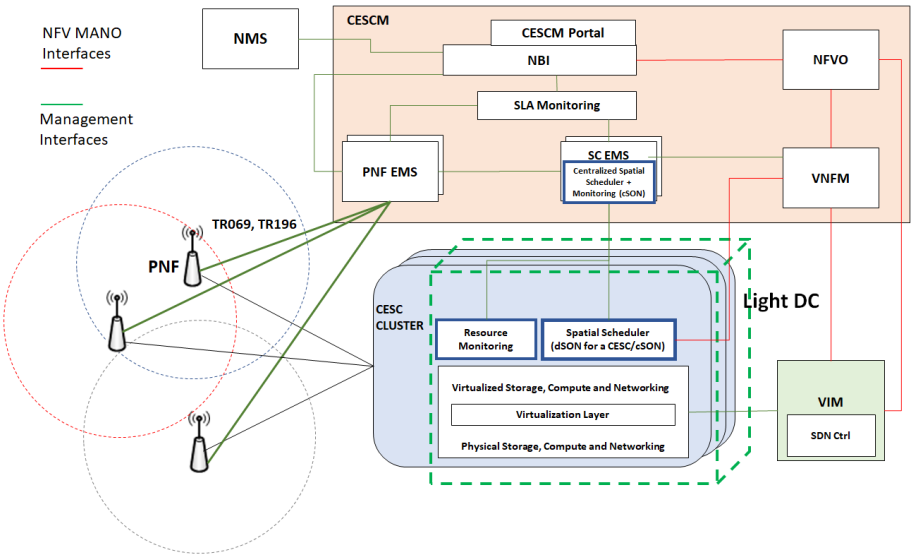

Figure 16: Mapping of centralized and distributed spatial scheduler in the SESAME system architecture [26]

3.1.4 Workflow of the Feature

As already discussed, the spatial scheduler can be used in either centralized (cSON) or distributed fashion (dSON). In centralized fashion, the problem solution makes the attempt to optimize the transmission of downlink packets for the users in each scheduling area relying on a global view of the network, alleviating at the same time the ICI problem caused by neighbouring downlink transmissions in the cluster. In this case, network statistics are aggregated considering the CESC cluster level situation. On the other hand, in distributed fashion the optimization is done based on local information. In this latter case, the optimisation is carried out for each scheduling area within a CESC or it could be even done for a group of them. More generally, the CESC(s) could be part of the slice of a tenant. The logical steps to run the Self-Optimising feature focusing on a reference CESC are as follows

1. For centralized spatial scheduler
    a. Find the number of active bearers in the whole CESC area.
    b. Acquire measurements done by all UEs in the CESC (Reference Signal Received Power [RSRP], Received Signal Strength Indication (RSSI)).
    c. Rely on the current MCS used in each bearer.
    d. Acquire the number of scheduling areas currently used (even one is possible in omni-directional mode).
    e. Solve the MC-GAP or the heuristic solution to find out whether an alternative way of partitioning the CESC coverage area is possible.
    f. Enforce the new configuration to the CESC users if this offers a better solution with respect to the previous settings; leave it unchanged in case the new global solution does not bring advantages.
2. For distributed spatial scheduler repeat the same steps as before but applied to each scheduling area of the reference CESC solving the MC-KP problem.

3.2 RAN Sharing

The SESAME architecture facilitates sharing the resources of the CESC both in the cloud infrastructure part (i.e., Light DC), and the radio access segment through RAN sharing, which leverages on the concept of resource slicing. In the scope of the 5G network, network services and tenant VSCNOs can be very dynamic in terms of the resource requests from the underlying infrastructure. This implies that the architecture must be able to provide the necessary features to accommodate such requests for existing and new virtual networks and services. Moreover, the network resources that are required to accommodate services must scale with the requests as defined by the SLA of the tenants. From a Self-Organisation (mainly in terms of self-configuration and Self-Optimisation) perspective, the scaling and reconfiguration of the resources that must be offered by the infrastructure can be done in an autonomous manner in reaction to network changes. Therefore, whenever specific conditions are met or alerts are raised, upon collecting information from the RAN, a Self-Organising network can automatically decide when and how to reconfigure the radio resources according to a specific policy. The Self-Organising mechanisms developed in SESAME ensure that network slices can be dynamically adjusted to meet the network/service requirements. This was thoroughly discussed in Section 2.2.

RAN sharing is an important part of this overall picture, which necessitates the capability to slice the radio resources available to the SESAME infrastructure into tenant-specific resource pools. Providing this capability in the SESAME architecture requires that the small-cells are dynamically (re)configurable and radio resources (Physical Resource Blocks in LTE) be dynamically sub-divided into non-overlapping chunks under the control of a dedicated RAN controller. For the purpose of demonstrating RAN sharing, the term RAN controller used here includes the necessary functions of the VIM to manage the RAN slices, as well as the interface to acquire statistics from the small cell/PNF and expose them to an SDN controller. To achieve this latter goal, the small cells must report statistics of the resource consumption (e.g., PRBs used in a slice) of the tenants/slices or within the whole cell to the controller and VIM. This additionally requires communication between the controller and the small cells, which can be carried out by a dedicated software module/agent. In case the small cell stack is virtualised, the software agent can also reside in the Light DC, whereas in case virtualisation is done at a higher level (i.e., S1 virtualisation), the software agent will be located directly in the PNF. From a Self-Organisation perspective, the controller implements the cSON logic that can dynamically reconfigure the RAN slices by issuing commands to the software agent. Furthermore, to ensure that slices do not overlap across multiple small cells in a way to avoid interference, the VIM keeps the global view of the resource allocation and utilisation across the RAN slices.

3.2.1 Implementation Details

As previously described, there are three main components that need to be developed to realize a Self-Organised RAN sharing feature: a centralized RAN controller, a reporting and reconfiguration mechanism exemplified by the software agent and the capability of the small cell to support slicing of the radio resources. For the sake of showing the implemented RAN sharing, the software agent is deployed at the PNF side. Figure 17 shows the prototype of the outlined RAN sharing concept and the overall architecture of the implementation. To elaborate on this, the 5G-EmPOWER controller takes the role of a VIM to manage the RAN slices (i.e., the RAN controller), and from the SESAME orchestrator perspective, 5G-EmPOWER embodies the functionalities of VIM and SDN controller that takes the responsibility of the RAN slices

(re)configuration. To complete the explanation, the "5G-EmPOWER eNB agent" is the software entity that carries out the commands from the 5G-EmPOWER controller until the actual reconfiguration of the resources. The RAN is implemented through the software-defined radio Ettus B210, and using the open source eNB software Open Air Interface (OAI). Core network functionality is provided by the Athonet's virtualised EPC.

An implementation of RAN sharing is shown in Figure 18, in which the waterfall diagram illustrates two slices in which one slide (on the left-hand side) is idle (no UE is actively connected) and another one (thick green coloured on the right-hand side) in which an active UE is receiving downlink traffic. The two slices of the RAN resources rely on two different schedulers. Specifically, the Best CQI scheduler is implemented in one slice, whereas the other relies on the simple round robin scheduler. It can be seen, that the active downlink uses a bandwidth of 2.4 MHz, or 13 PRBs.

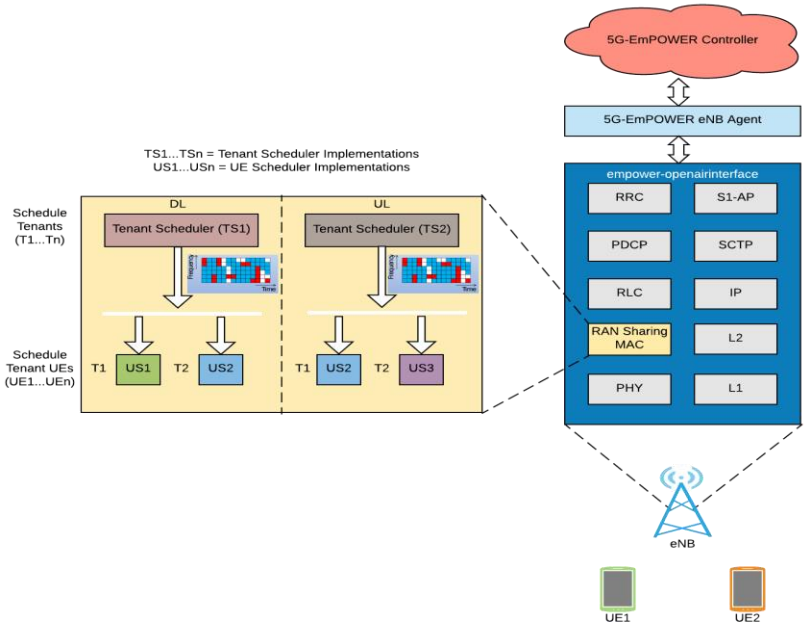

Figure 17: RAN sharing in small cells.

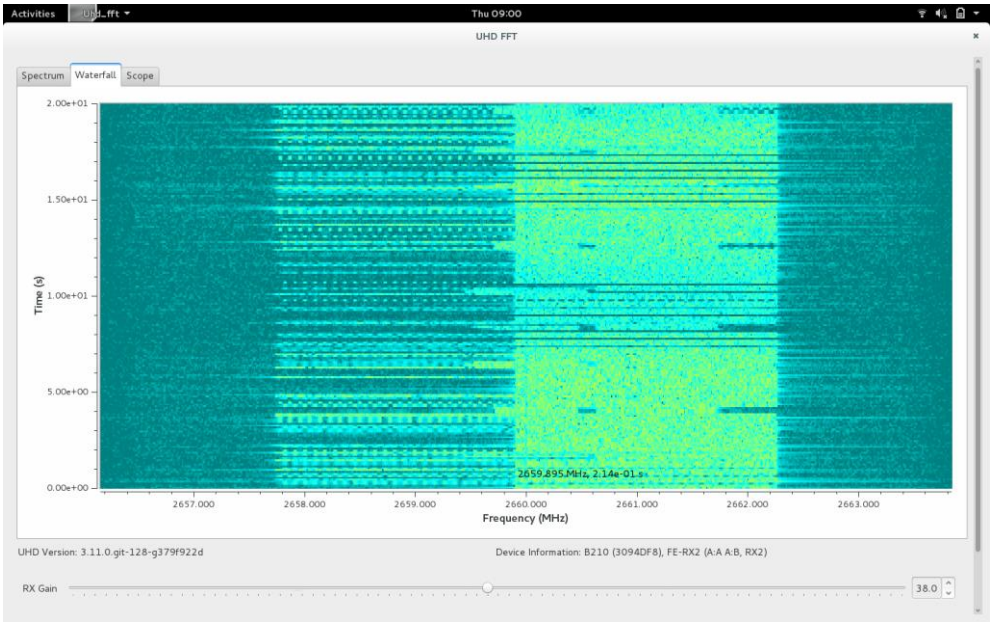

Figure 18: Screenshot of the downlink traffic addressed to a UE that belongs to one tenant (half right-hand side of the diagram). Y axis presents time in seconds (0 to 20s) and x axis shows frequencies in MHz (~ 2656MHz – 2664MHz).

3.2.2 Mapping to the SESAME components

From the SESAME perspective, the RAN sharing feature has direct mapping to the architecture including the CESCM as shown in Figure 19. The 5G-EmPOWER controller has direct mapping to the VIM and SDN controller in the SESAME architecture to manage the RAN slicing process, providing also RAN sharing specific control interfaces that enable the communication and reconfiguration of the small cells. The cSON feature needed to realize the dynamic reconfiguration of the RAN slices and individual slice's scheduling is the central intelligence that makes decisions for the different entities within the SESAME architecture. In the full-fledged RAN sharing feature, the 5G-EmPOWER controller is the VIM, that interacts with the CESCM (e.g., orchestrator) to create, manage, and delete RAN slices within the CESC cluster. The 5G-EmPOWER agent is a cell-tied feature and therefore can be deployed in the small cells/PNFs for configuring the RAN slices. Alternatively, in case part of the small cells stack is virtualised (functional split lower then S1 virtualisation), the agent could also be deployed in the micro-servers composing the Light DC cloud environment.

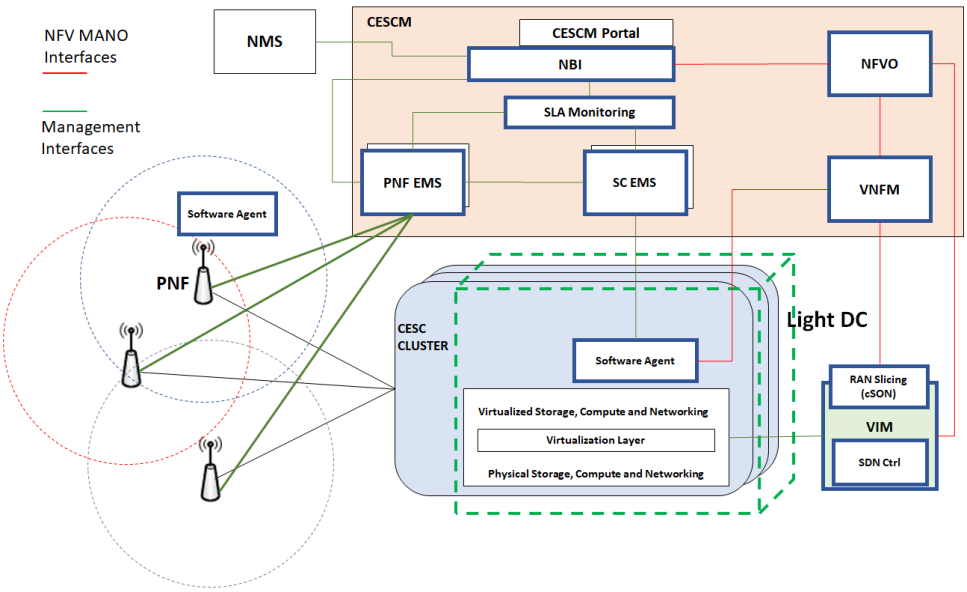

Figure 19: Mapping of RAN sharing Self-Organising feature in the SESAME architecture [26]

3.2.3 Workflow of the feature

Two distinct workflows have to be described for realizing RAN sharing within the SESAME architecture, exemplified by the prototype implementation shown above. First, the request for the creation of a slice is based on the observation done by the RAN controller that an operator (or service provider) requires additional resources within the CESC cluster. Alternatively, it could also be an explicit request of a tenant through the CESCM portal or a dedicated API, in which case the request is communicated to the 5G-EmPOWER through its northbound RESTful APIs. The request must specify the attributes of a slice (i.e., in a template carrying the number of PRBs, the type of scheduler to use, the sharing agreement with other operators, etc.) based on the specific SLA of a tenant, and it must also specify in which CESC (or a group of them) the slice has to be deployed. The information is used by the RAN controller to create the slice through the software agent. The second workflow consists in the reconfiguration of a tenant slice relying on the software agent. In both cases, the Self-Organisation is based on a bottom-up communication flow in which the software agent communicates the state of resource utilization to the 5G-EmPOWER controller where the decision-making process is done.

As a last consideration, given the nature of the RAN sharing function, the intelligence that makes the decision to create, update and delete a slice can effectively reside at different places of the SESAME architecture. The implementation shown above puts the intelligence at the 5G-EmPOWER controller side to enable the cSON approach but other choices are also feasible. For instance, upon doing the functional split of the small cell protocol stack, the cSON function can be placed in the CESCM.

3.3 Edge Caching in SESAME

The SESAME Light DC architecture provides the fundamental support to bring multi-access edge computing (MEC) benefits to the end-users in a transparent and Self-Organised manner [3]. More specifically, the Light DC provides the edge-cloud environment that can support small cell

and service functions to optimize content delivery to the end-users. One of the challenges consists of ensuring low end-to-end latency for applications and services with specific quality-of-service (QoS) requirements. This can be a bottleneck for many applications in traditional network architectures, which can be overcome in 5G. A solution consists of exploiting the SESAME architecture to cache popular content in the Light DC. This is obviously the case of Over-The-Top (OTT) service providers that want to reach their end customers, which can be done directly acquiring a slice of resources within the Light DC or otherwise leasing resources from a VSCNO.

To alleviate latency concerns for many 5G network applications, an edge caching framework that leverages on network softwarization is presented, which enables tenant networks/VSCNOs to realize mobile edge-computing benefits and bring services closer to the end users. Through the SESAME system, the caches can be realized as VNFs orchestrated inside the Light DC. The key idea is to automatically create, modify and delete a cache, as well as store/remove contents from the cache relying on a level of intelligence that is based on content popularity and cache missed rate ([4], [5]). The decision model suits particularly well the case of video. In this way, contents can be managed on a per tenant basis leading to improved joint management of the CESC cluster and backhaul resources (reduced latency, improved management of storage and reduced backhaul traffic). As such, the decision-making process for caching content at the edge is modelled as a dSON function.

3.3.1 Implementation Details

MEC caching has been implemented through modular software blocks within one CESC. Figure 20 shows the high-level architecture of the edge caching. This architecture is based on a few fundamental components that are briefly illustrated. Looking at Figure 20 from the top, a MEC application can interact with the MEC Framework to provide a service to the end users with improved quality of service. The Hosting infrastructure refers to the hardware and software combination that provide the edge platform to support the MEC server and the edge-resident service applications. The MEC Application Platform services provide the fundamental middleware functionalities to support several MEC applications inside the SESAME Light DC architecture. The MEC-RAN Information Interface is a RAN-specific interface between the MEC server and the underlying physical (or virtual) network. It provides a real-time insight into the radio network information and user location awareness.

Relying on the possibility to cache contents in an autonomous manner relying on a level of intelligence to optimise resource utilisation (i.e., Self-Organizing manner), this can be done as shown in [4] and [5], based on the content popularity (e.g., number of content requests), which can be exposed from the MEC server to the dSON algorithm.

In order to prove the effectiveness of deploying edge caching, the round-trip time (RTT) was measured with and without caching as shown in Figure 21. The RTT was not measured for a video content but rather measured as the latency incurred to connect to the homepage of the websites shown on the horizontal axis. This was done to avoid the problem of buffering for video packets. The figure clearly highlights the benefits of caching.

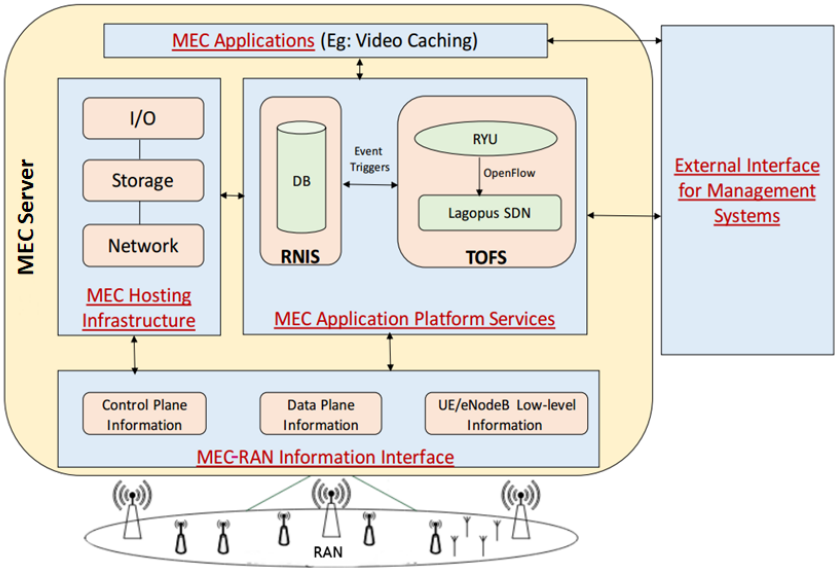

Figure 20: Level Software-Defined architecture for edge-caching

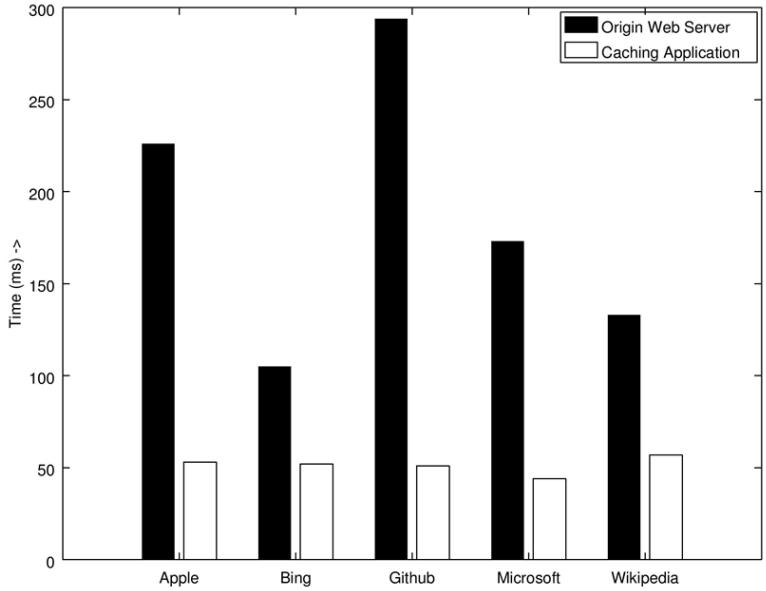

Figure 21: Measured RTT for edge caching. The black bars denote the latency measured to connect to the web pages indicated on the horizontal axis. The white bars report the latency in case the web page content was cached

3.3.2 Mapping to the SESAME components

In general, the edge caching feature has a direct mapping in the SESAME architecture as shown in Figure 22. The fact that the different components of this feature can be deployed within the Light DC offer the advantage to orchestrate them as VNFs, although this latter part is not fully developed in the current implementation. The MEC applications are meant to be located at the network edge to provide low latency, high QoS services and therefore the logical host is the SESAME Light DC infrastructure. The MEC framework provides the core intelligence of the edge

service. While the existing implementation places the controller alongside the rest of the framework components, this is not a strict requirement for all MEC applications. In Figure 22 the blue edge boxes indicate the SESAME entities that can be involved in the overall configuration and life-cycle management of edge caching (or a MEC service in general) including the Self-Organisation part. The respective EMS per each VNF monitors the health of the MEC application and it reports to the SLA monitoring thus facilitating a dSON approach for the placement and configuration of the caches. Relying on SESAME terminology, it is understood that MEC caching does not only involve SC functions but service VNFs as well, for which the MEC applications provide an exemplary case. Therefore, the full-fledged Self-Organising MEC caching feature shall be deployed through the MEC Framework and Application in CESC or a group of them (involving the Light DC in this case) as virtualised network functions and the respective placement, configuration and life-cycle involve the different components of the SESAME architecture.

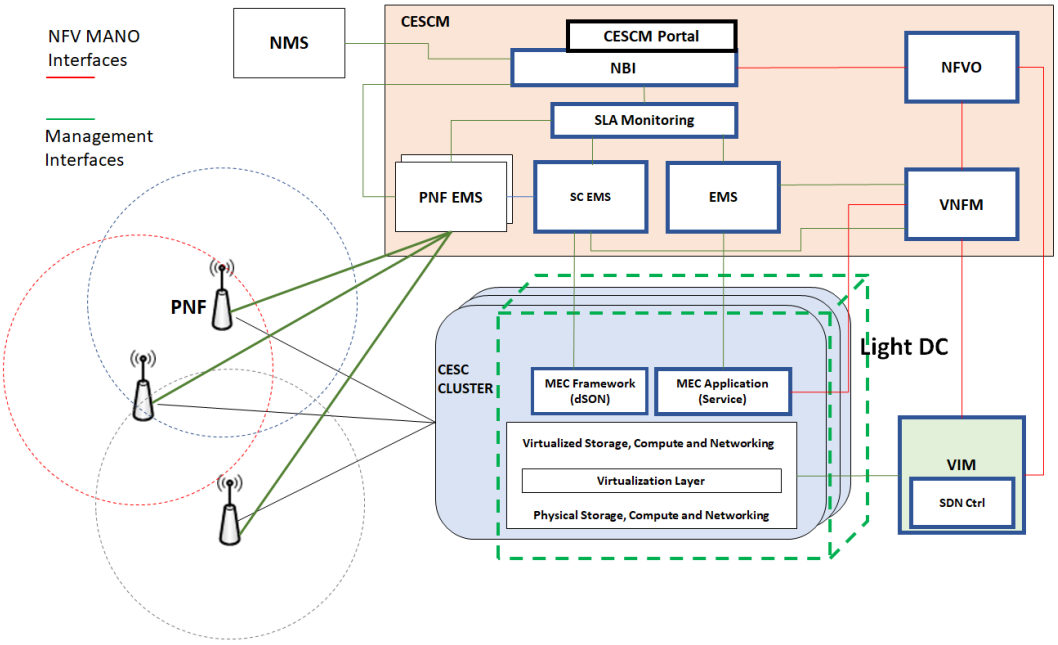

Figure 22: Mapping of MEC Application framework onto the SESAME architecture [26]

3.3.3 Workflow of the feature

To support MEC applications, the MEC framework has to collect different types of information related to the RAN, as well as to the specific service. Figure 23 shows a high-level overview of the traffic flows that the MEC Framework supports for an edge content caching application. The traffic flows in the MEC server are handled in Pass-through mode where the end-user service is provided via the core network (red dashed line) and in End-point mode where the end-user receives the service from the MEC application placed inside the CESC (orange dotted line). It is worth to stress that to perform either mode the GTP traffic is intercepted, de-capsulated first and re-encapsulated afterward. This is the implementation of the S1-level virtualisation included in SESAME.

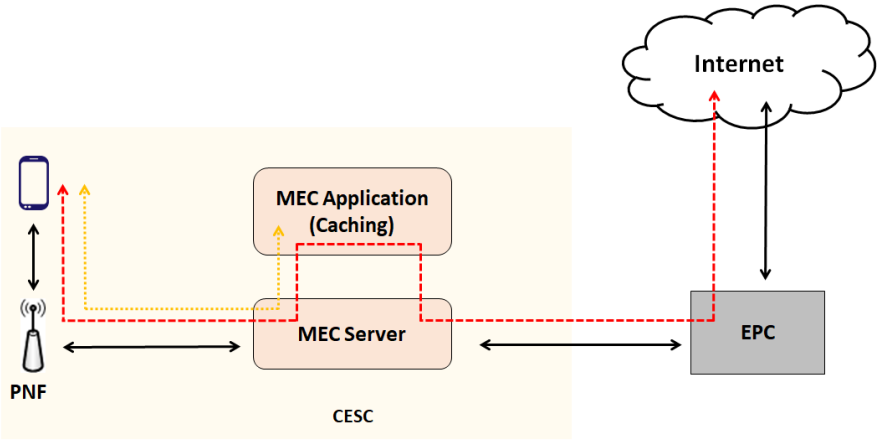

Figure 23: A high level overview of traffic flows in edge caching

3.4 Transmit Frequency and Power Selection

Transmit Parameter Selection is a SON feature that enables a cell to automatically select a number of its key operating radio parameters such as:

- Transmit Frequency selection,
- Transmit Power setting,
- PRACH root sequence selection,
- PCI selection.

3.4.1 Implementation Details

The following subsections provide the details on the implementation for

3.4.1.1 Transmit Frequency Selection

When supported, this feature enables a cell to select its transmit frequency from a list of candidate frequencies. This selection is made following a radio environment scan (also known as Network Listen or NWL) and is driven by the desire to minimize or eliminate interference from and with neighbouring cells.

When implemented as a distributed SON (dSON) feature, each cell will typically select a frequency with the lowest received signal levels from neighbouring cells. In a centralised (cSON) or hybrid (hSON) implementation, some high-level entity will automatically configure a subset of cells using traditional cellular network planning methods. This may, *for example*, be used to minimise interference between a cluster of small cells and the macro network.

Due to the wideband nature of LTE carriers (up to 20 MHz compared to 5 MHz for 3G) automatic transmit frequency selection may not be a very useful feature. This is especially where an operator is engaged in a frequency re-use of one and configures each cell to utilise their entire spectrum allocation.

This sub-feature is not considered in SESAME.

3.4.1.2 Transmit Power Setting

This feature enables a cell to select its transmit power from within a defined range. Depending upon the implementation, the goal of the selection algorithm may be to minimise interference with neighbouring cells or to achieve a certain minimum coverage. Either way, the goal is to select the minimum transmit power that meets the desired target.

In a dSON implementation, each cell separately uses the results of radio environment scan to select a TX power level based on detected neighbour cells. In a cSON or hSON implementation, a higher level entity determines the TX power for a set of cells, giving consideration to factors such as:

- Desired coverage,
- Interference between the cells under consideration,
- Interference between the cells under consideration and other neighbour cells.

In SESAME, the PNF provided by the ip.access E40 provides this capability and operates in dSON mode. Following a radio environment scan, it attempts to set its transmit power in order to achieve a specific coverage target as defined by two parameters:

- **FAP Coverage Target** specifies the target path loss for the cell, measured in dB,
- **Cell Edge SINR Target** specifies the desired signal quality at the cell edge, specified as a Signal to Interference Ratio.

The PNF may utilise one of two algorithms as specified by configuration:

Algorithm 1 – Open Loop Mode

In open loop mode, the transmit power of the cell, as defined by its Reference Signal Power parameter, is set according to the following formula:

$$RSP = TNPRE + CESINR + PL_{Target} + Fade_{Margin}$$

where:

RSP	is Reference Signal Power, specified in dBm,
TNPRE	is Thermal Noise Per Resource Element (15 kHz), assuming an 8 dB figure for the UE (see below)
CESINR	is the configured Cell Edge SINR target (default value 3:1),
PL_{Target}	is the configured target path loss, in dB,
$Fade_{Margin}$	is the configured fade margin, in dB (default 8).

$$RSSI_{thermal}$$
$$= Thermal\ noise\ at\ UE\ in\ Watts\ assuming\ 8\ dB\ noise\ figure\ for\ cell\ A's\ BW$$
$$= -174 + 10 * log10(\ cell\ A's\ BW\ in\ Hz) + 8\ dB = -174 + 41.76 + 8 = -124.24$$

Algorithm 2– Closed Loop Mode

When the SESAME PNF is operating in closed loop mode, the transmit power of the cell, as defined by its Reference Signal Power, is set to achieve the configured Cell Edge SINR, taking into account the downlink interference from neighbour cells and computing a path loss bound up to which the target SINR is possible.

This is achieved as following a radio environment scan as follows:

- The path loss (PL_{Ai}) to each co-channel neighbour in the "REM[5]" list is calculated,
- A path loss bound, required to achieve a stable cell edge SINR is computed by considering the path loss to each detected co-channel neighbour,
- A Rise Over Neighbours (RON) value is then computed. This is the RSSI contribution from non-neighbours over known neighbours.
- The RON value is then deducted from the previously computed path loss bound to arrive at a final path loss bound ($\widehat{PL_A}$).
- In co-channel deployments, the value used by the PNF is then the minimum of ($\widehat{PL_A}$) and the configured coverage target, in dB.
- In non co-channel deployments, the coverage is reduced by the detected RON in order to achieve stability.

The above algorithm was derived following investigation using an LTE simulator to assess the impact of the different algorithms and settings on the resulting SINR and CQI maps, together with data rates and data rate densities as figures of merit for the different algorithms. The simulator includes a model of NWL which is used to measure the surrounding cells and provide inputs to the algorithms. The resulting algorithm is summarised by Figure 24 below:

[5] Radio Environment Management. This list is populated by the Automatic Neighbour Relations function.

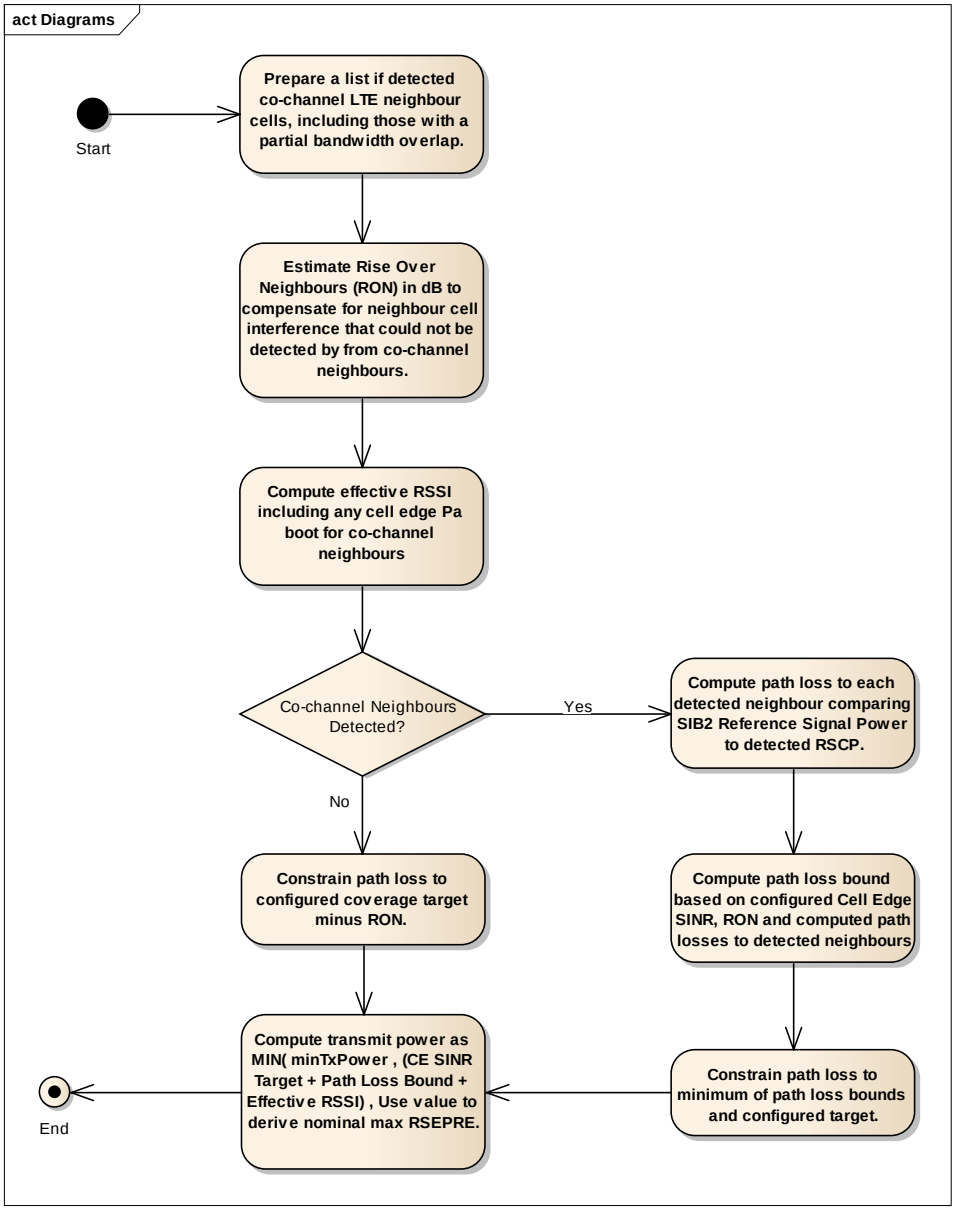

Figure 24: Closed Loop TX Power Algorithm

3.4.1.3 PRACH Root Sequence Selection

The Root Sequence Index is broadcast as part of PRACH configuration in SIB 2[6] of an LTE cell. It allows a UE to calculate which PRACH preamble to use to attach to the cell. Neighbouring cells are configured with different Root Sequence indexes in order to ensure that UEs performing a random access attempt are recognised by the correct cell.

The SESAME PNF supports the selection of PRACH Root Sequence Indexes from a configured list or permitted values. It does this by eliminating all contiguous indices of detected co-channel LTE neighbours from the configured list and then selecting the required values from those

[6] For more details see: http://lteinwireless.blogspot.gr/2012/12/sib2-in-lte.html

remaining. In the unlikely event[7] that there are no indexes remaining, the PNF flags an error and does not enter service.

3.4.1.4 PCI Selection

The SESAME PNF supports the selection of its Physical Cell Identity from a configured list of permitted values. It does this by eliminating the values of detected co-channel LTE neighbours from the configure list to derive a Permitted PCI List. Each entry in this Permitted PCI List has three associated counts: PCI Mod 3 count, PCI Mod 6 count and PCI Mod 30 count. The PNF iterates through the list of co-channel LTE neighbours detected by the ANR function and where PCI Mod *X* of the neighbour cell equals PCI Mod *X* of an entry in the Permitted PCI List, the associated count is incremented. The PNF then selects a PCI as follows:

- If any configured PCI has a zero Mod 3 count, the first such value encountered is selected,
- If no PCIs have a zero Mod 3 count then the first PCI with a zero Mod 6 count is selected,
- If no PCIs have a zero Mod 3 or Mod 6 count then the first PCI with a zero Mod 30 count is selected.

In the unlikely event of all Mod 30 PCIs have a non-zero count, the PNF selects the value with the lowest Mod 30 count.

3.4.2 Mapping to the SESAME components

Currently, TX Parameter Selection is implemented entirely in the PNF and is performed following a radio environment scan. The SCNO uses the EMS to configure the PNF and specifies values that include:

- The Transmit frequency of the PNF,
- The TX power algorithm (open or closed loop),
- The coverage target (in dB).
- The target cell edge SINR,
- A list of permitted PRACH root values,
- A list of permitted PCI values,
- The desired neighbour cell ranking strategy,
- An optional list of DL E/U/ARFCNs to scan for each supported RAT. If a list is empty, the PNF scans all of each configured band.

3.4.3 Workflow of the feature

The scheduling of the radio environment scan is controlled by standard parameters in the [TR196] data model and is configured by the EMS.

The overall workflow is:

- Perform the environment scan,
- Record the details of detected neighbours,

[7] The default range provides a total of 838 possible indices.

- Perform TX parameter detection,
- Report detected neighbour and selected parameter values to the EMS.

Note that when PNFs are deployed in close proximity, care needs to be taken with their configuration. If the PNF does not have a dedicated radio for environment scanning, it will need to be taken out of service in order to use its radio for this purpose. If PNFs in the same area are taken out of service to perform an environment scan at the same time then they will return false results. To avoid this problem, the PNFs are either configured with explicitly staggered scan times or have some degree of randomisation incorporated into their scan schedule.

3.5 Automatic Neighbour Relations (ANR)

Automatic Neighbour Relations (ANR) is a Self-X feature where a cell uses a periodic radio environment scan to discover neighbour cells that are candidates for hand-out. Details relating to these neighbours are:

- Stored by the cell and reported to the EMS,
- Broadcast in SIBs 4, 5, 6 and 7 for use by UEs,
- Used to assist with the measurement configuration of UEs,
- Used to influence the handling of measurement reports received from UEs.

In addition to this period scan, the SESAME PNF supplements this information by requesting UEs to perform measurements. As the PNF may not be located in a position that is representative of the UEs that it serves, UE measurements allow it to improve the quality of its decision making with respect to which neighbours are the best candidates for handover.

In SESAME, this behaviour is fine-tuned to take account of the fact that the PNF may be serving multiple virtual cells, each with different handover needs.

3.5.1 Implementation Details

Standard [TR196] data model parameters determine the time of day and repeat period for radio environment scans. Additional parameters may be used to constrain the range of frequencies to scan in terms of specific ARFN, UARFCN or EARFCNs as it is useful for operators to speed up the process my omitting frequencies that they are not using.

For each detected neighbour, the PNF populates parameters in the [TR196] data model with details of the neighbour. For example, for an LTE neighbour, this includes but is not limited to:

- The identity of the neighbour cell (e.g., PLMN and Cell Identity in LTE),
- The list of PLMNs supported by the neighbour,
- Physical cell ID,
- EUTRAN carrier frequency,
- Received RSRP level, specified in dBm,
- Received RSRQ level, specified in dB,
- Received Signal Strength Indicator (RSSI), specified in dBm,
- Tracking Area Code (TAC).

These neighbours are then ranked according to the configured strategy which is one of:

- Disabled – no ranking is performed,
- RSRP Only – neighbours are ranked by Reference Signal Received Power only,
- RSRQ Only – neighbours are ranked by Reference Signal Received Quality only ,
- RSRP and RSRQ - neighbours are ranked by a combination of Received Signal Code Power (RSCP) and RSRQ,
- UE Measurements.

When the ranking strategy is set to UE Measurements, all detected neighbours are initially assigned the same rank. Once the PNF has entered service, a configurable path loss parameter is used to identify UEs at -or near- the cell edge. After expiry of an idle time, periodic measurement reports are enabled on a sub-set of UEs deemed to be at the cell edge. These measurement reports are then used to adjust the rankings of neighbours in the list; those that are detected most often are moved to the top of the list. In addition, neighbours not detected by the radio environment scan may be added to the bottom of the list. After the expiration of a configurable time during which no new neighbour cells are detected, periodic measurement reports are disabled again.

For a non- Multi Operator Core Network (MOCN) supporting cell, neighbours with the same PLMN as the scanning cell are normally ranked according to the configured strategy and broadcast, in order, such that the "best" neighbour appears first in the list for the appropriate System Information Block (SIB). Where the number of detected neighbours is very large (for example in a dense small cell deployment) it may not be possible to broadcast the details of every neighbour and, thus, only broadcasting the details of the N best ones makes perfect sense.

In SESAME, the PNF supports multiple PLMNs so this behaviour is modified to provide equal ranking to neighbours supporting each of the PLMNs served by the PNF. For example, consider the following neighbours ranked in absolute order:

1. PLMN A Cell #1A,
2. PLMN A Cell #2A,
3. PLMN A Cell #3A,
4. PLMB B Cell #1B,
5. PLMN A Cell #4A,
6. PLMN B Cell #2B,
7. PLMN C Cell #1C,
8. PLMB B Cell #3B,

A SESAME PNF supporting all three PLMNs might broadcast these neighbours in the following order:

1. PLMN A Cell #1A,
2. PLMB B Cell #1B,
3. PLMN C Cell #1C,
4. PLMN A Cell #2A,
5. PLMN B Cell #2B,
6. PLMN A Cell #3A,
7. PLMN B Cell #3B,
8. PLMN A Cell #4A,

Thus, neighbour cells #1A, #1B and #1C are given more-or-less equal ranking even though #1B and #1C have a significantly lower absolute rank order than cell #1A.

Similarly, when a UE accesses the PNF using the RRC connection process, the SESAME PNF uses the UE's *Selected PLMN* to identify those neighbours that support this PLMN and sends it a tailored Measurement Configuration so that it only performs measurements on those frequencies that support neighbours to which it can hand out.

3.5.2 Mapping to the SESAME components

This feature is implemented primarily in the PNF. It uses information from the EMS and each VSCNO to determine:

- The PLMNs supported by the PNF (as part of the Virtual Cell provisioning process);
- When and how often to perform a radio environment scan;
- The frequencies on which to perform a radio environment scan. Currently, in the SESAME PoC, this is configured by the SCNO but it is envisaged that a commercial deployment might allow each VSCNO to specify their own neighbour frequency details,
- The neighbour cell ranking strategy to employ;
- If the "UE measurements" strategy is used, a path loss boundary at which measurements may be triggered.

3.5.3 Workflow of the feature

The scheduling of the radio environment scan is controlled by standard parameters in the [TR196] data model and is configured by the EMS.

The overall workflow is:

- Perform the environment scan,
- Record the details of detected neighbours,
- Rank neighbour cells according to the configured strategy and report to EMS via [TR196] "in use" parameters,
- If the ranking strategy is set to "UE Measurements" then the PNF enables periodic measurement reporting on a sub-set of idle cell edge UEs to adjust rankings and add newly detected cells.
- After a period of no new information being detected, periodic measurement reporting is disabled again.

3.6 Self-Optimisation of Admission Control (AC) in Multi-Tenant Networks

A new multi-tenant Admission Control (AC) approach for RRM was proposed in section 5 of Deliverable D3.2 [1] to regulate the share of radio resources in each cell among different tenants by deciding on the acceptance/rejection of the E-RABs of each tenant. The proposed approach takes into account the amount of capacity that is committed to each tenant, as specified in the SLA terms.

To adapt to different network conditions and to further optimise the performance of the AC process, a Self-X function is used to dynamically modify the AC parameters. This Self-X function

intends to achieve an efficient utilisation of the available radio resources by exploiting traffic multiplexing principles at both cell and multi-cell level. This Self-X function is framed within the use case "Optimized Radio Network Capacity Planning and Operation mechanisms of the Small Cell Network Operator" defined in deliverable D2.1 [28].

Figure 25 illustrates the functional architecture of the proposed framework for AC Self-Optimisation. The AC function at the *n*-th cell is triggered from each tenant's Evolved Packet Core (EPC), i.e., from the Mobility Management Entity (MME), indicating the bit rate R_{req} to be guaranteed to the E-RAB. The Self-X function is a hybrid SON approach in which a distributed SON function (dSON) running at each cell optimizes some parameters with the support of a centralized SON (cSON) function running at the Element Management System (EMS) that can make decisions based on an overall view encompassing multiple cells

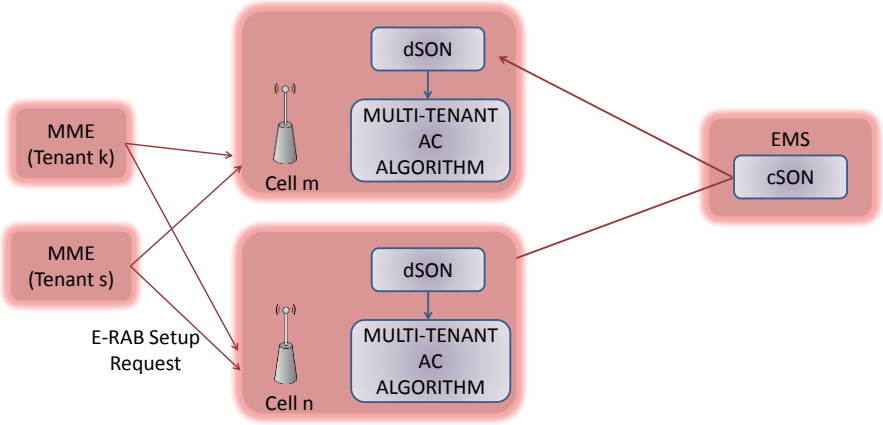

Figure 25: Functional architecture of the AC Self-Optimisation

The operation of the AC function is based on two different checks performed at each cell (see [1], [29] for details):

1) Capacity check at cell-level: This capacity check evaluates the aggregated number of Physical PRBs used by all the tenants in a cell after accepting the E-RAB request and ensures that the cell has sufficient physical resources for serving the new E-RAB.

2) Per-tenant capacity share check: This check establishes an upper bound in the PRBs used by the E-RABs of a tenant in accordance with the capacity contracted through the SLA, which for the purpose of this work is defined by the so-called Scenario Aggregated Guaranteed Bit Rate (SAGBR) that establishes the total bit rate to be guaranteed for all the E-RABs of a tenant across all the deployed cells.

Based on the SAGBR, the per-tenant capacity share check establishes a bound in the fraction of PRBs used by the s-th tenant in the n-th cell given by $C(s)+\Delta C(s,n)$, where:

- $C(s)$ is the nominal capacity share, defined as the ratio between the SAGBR(s) of the *s*-th tenant with respect to the aggregated SAGBR of all the tenants.
- $\Delta C(s,n)$ is a term that enables some flexibility on the capacity share per tenant to account for unused capacity left by other tenants and to cope with heterogeneities in the spatial traffic distribution.

The hybrid SON function is in charge of dynamically adjusting the term $\Delta C(s,n)$ of the per-tenant capacity share check. As explained in section 5.1.2 of [1], this computation is done by

decentralized component (dSON) running at the *n*-th small cell. The computation of $\Delta C(s,n)$ combines the following terms (see [1], [29] for details):

- A term $\Delta C_e(s,n)$ that measures the extra capacity share that is potentially available for the *s*-th tenant in the *n*-th cell when the other tenants $s' \neq s$ are leaving unused capacity. This term is computed by the dSON component.
- A term $\Delta C_b(s,n)$ that measures the increase or decrease in the capacity share to should be applied in the *n*-th cell to ensure that the average PRB utilisation of the *s*-th tenant across all the cells equals $C(s)$. This term is computed by the cSON component.

3.6.1 Implementation Details

The proposed AC Self-Optimisation process involves three main components, namely, the AC function that performs the two abovementioned capacity checks, the dSON function which computes the term $\Delta C(s,n)$ and the cSON function that computes the term $\Delta C_b(s,n)$ and provides it to dSON.

Both the AC and the dSON functions run at the CESC, so in principle, they can be implemented either as a PNF or as a VNF. The implementation of RRM/SON functions as VNFs has different potential benefits, such as: (i) scalability and flexibility in how RRM/SON functions are deployed within a CESC cluster, thanks to the capability of adding/removing the services as required, where and when needed and to upgrade them without hardware/firmware changes of the radio equipment; (ii) ease of integration of RRM/SON solutions from multiple vendors if common interfaces are adopted, which facilitates a more open market where third-parties could provide RRM/SON components to be flexibly plugged into the network; (iii) higher efficiency in power consumption by centralizing functions in some micro servers of the Light DC; (iv) more flexibility in accommodating variable computation requirements of the RRM/SON functions by scaling up/down the amount of resources allocated to them across the LightDC micro servers.

Based on the above considerations, this section explores the possibility of implementing the proposed AC/dSON functions as a VNF. As for the cSON component, it runs at the EMS of the CESCM, so the possibility of implementing it as a VNF is not considered in SESAME, since the CESCM runs outside the Light DC.

The following sub-sections elaborate on the required architecture for implementing this function in relation to the existing SESAME components.

3.6.2 Mapping to the SESAME components

Figure 26 illustrates a particularisation of the SESAME architecture of [30] for implementing the abovementioned AC Self-Optimisation process. As previously discussed, the cSON component runs at the EMS, while the AC and dSON components are implemented as a single VNF running on the virtualized execution platform provided by the Light DC.

The implementation of RRM/SON functions as VNFs, like the case of AC/dSON considered here, requires that radio network related information available within the small cells be accessed by these functions. For that purpose, the implementation considered here assumes another VNF, denoted in Figure 26 as Radio Network Information Service (RNIS), which handles the acquisition, processing and collection of different measurements to monitor the status of the RAN and exposes them to the AC/dSON function. Indeed, this functionality has been placed separated from the AC/dSON function, because it is envisaged to be a common VNF that can

also provide measurements to other RRM/SON functions. This will avoid separate access to the same information from multiple deployed RRM/SON functions The exposed information by the RAN could contain a wide range of metrics such as measurements and statistics information related to the user plane, information related to UEs served by the cells, cell load, throughput measurements, etc., delivered at the relevant granularity (e.g., per UE, per cell, per period of time). Moreover, the radio network information could be profiled according to the specific needs of each RRM/SON function.

The measurements to be collected by the AC/dSON function are Layer 2 measurements [31]. Therefore, within the prototyping framework in SESAME project, which assumes a functional split at the S1 interface so that the SC PNF includes the whole radio protocol stack, these measurements are performed within the SC PNF. Therefore, these measurements can be obtained in SESAME from the SC PNF by means of Performance Management (PM) reports in the form of eXtensible Markup Language (XML) files including the measurements specified in [32] through a management interface such as the TR-069 [33].

While the interaction between the RNIS VNF and the SC PNF could be conditioned by vendor specific implementation of the SC PNF management interface, the approach for the RNIS VNF to expose the radio network information is the adoption of open interfaces, in line with the on-going work in European Telecommunications Standards Institute (ETSI) about the RNIS defined in [34] for MEC applications. This is the reason for adopting the terminology RNIS for the VNF considered here.

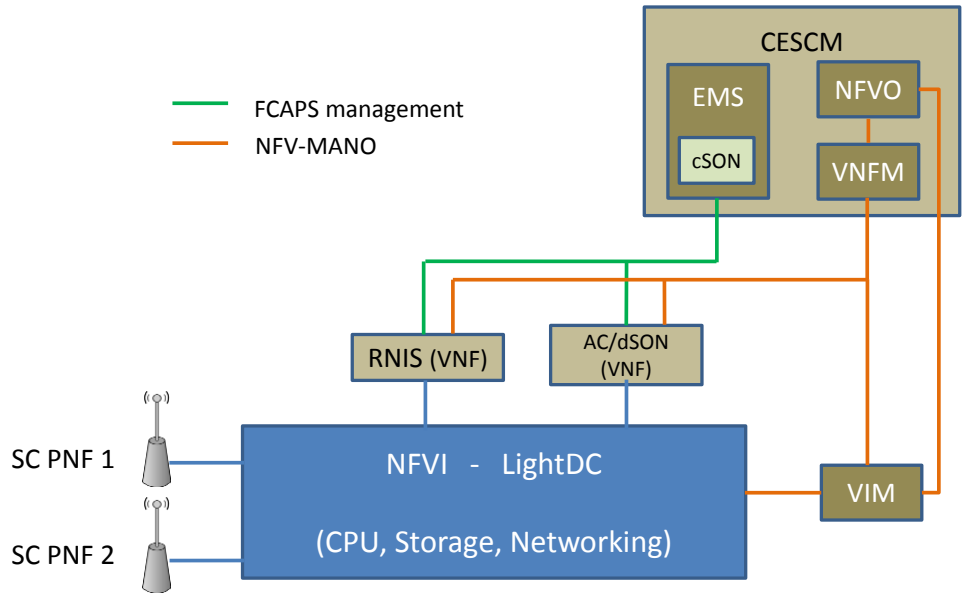

Figure 26: Implementation of the AC Self-Optimisation function in the SESAME architectural framework.

As depicted in Figure 26, the management of the RNIS VNF and the AC/dSON VNFs involves two different frameworks, namely the Fault, Configuration, Accounting, Performance and Security (FCAPS) management and the NFV-MANO.

The FCAPS management, whose interactions are depicted in green in Figure 26, is performed by the EMS and includes a set of functions for configuration and re-configuration of the operational parameters of the RNIS VNF and the different RRM/SON functions or for the collection of the PM measurements characterizing the operation of these functions. For the case of AC Self-Optimisation considered here, the EMS adjusts the term $\Delta C_b(s,n)$ and provides it to the AC/dSON

VNF. This is based on analyzing the PM measurements of multiple cells, which can be obtained either directly from each SC PNF or through the RNIS VNF.

The NFV-MANO, whose interactions are depicted in orange in Figure 26, involves the management functionalities provided by the NFVO, VNFM and VIM to support the lifecycle management of the RNIS and AC/dSON VNFs (e.g., instantiation, scaling, termination) as well as of the Network Services (NS) in which these VNFs are chained as components across the Light DC.

3.6.3 Workflow of the feature

Based on the algorithmic solution presented in section 5 of [1], the AC/dSON VNF requires the following input data:

- Average number of PRBs that the Packet Scheduling of the *n*-th cell has assigned to the E-RABs of each tenant. This is used in order to perform both the capacity check at cell level and the per-tenant capacity share check. This measurement can be obtained by extending the "PRB usage for traffic" measurement of [32] to include per-tenant measurements. The average of this measurement should be done during a time window *T* that will be typically in the order of some tens of seconds. This measurement will be provided by the RNIS VNF.
- Estimation of the bit rate per PRB that can be obtained in the n-th cell. This estimation can be obtained by dividing the aggregate IP throughput of the cell from [32] by the aggregate PRBs that have been allocated for transmission. The computation should be done during a certain time window T_e that will be typically in the order of some tens of seconds. The computation of this measurement will be done by the RNIS VNF, which will deliver it to the AC/dSON.
- Term $\Delta C_e(s,n)$. This term will be computed by the cSON, which will deliver it to the AC/dSON.
- Bit rate requirements of the new E-RAB requesting admission. This information is included in the "E-RAB Setup Request" and the "Initial Context Setup" messages that are originated at the MME to trigger the AC process of a new E-RAB.
- Nominal capacity share C(s) of each tenant. This information will be provided by the EMS when configuring the AC/dSON VNF function.

As for the cSON function, it requires the average number of PRBs assigned to the E-RABs of each tenant. This information has to be collected from the different cells, and can be obtained by extending the "PRB usage for traffic" measurement of [32] to include per-tenant measurements. This information can be provided by the RNIS VNF or can be extracted from the PM reports provided by the SC PNF.

Figure 27 depicts the signalling workflow that illustrates the interactions between the different components of the AC Self-Optimisation process once the different VNFs have been instantiated and are operative. It is explained in the following:

1) The cSON at the EMS continuously collects measurements of the PRB usage of the different tenants from each of the CESCs in the scenario. Measurements are collected with a relatively long-term periodicity involving some tens of seconds or some minutes. Collected measurements are averaged in a time window W_{cSON}, typically set to a few hundreds of seconds. Based on this, the cSON computes the term $\Delta C_b(s,n)$ to ensure that the average PRB utilisation of the *s*-th tenant across all the cells equals $C(s)$ (see expression (19) in [1]).

2) Then, the cSON delivers to CESC 1 the computed value $\Delta C_b(s,n=1)$ for all s=1,..., S tenants.

3) The AC/dSON function of CESC 1 collects the measurements of the PRB usage of each tenant in that CESC, as well the estimated bit rate per PRB. The measurements are stored and averaged with the configured time windows (i.e., T for applying the two checks of the AC process and W_{dSON} for determining the $\Delta C_e(s,n)$ term).

4) Based on the collected measurements the dSON function computes the $\Delta C_e(s,n)$ term using the expression (17) of [1]. In addition, based on the information received from the cSON, it computes as well the term $\Delta C(s,n)$ of the per-tenant capacity share check following the expression (16) in [1].

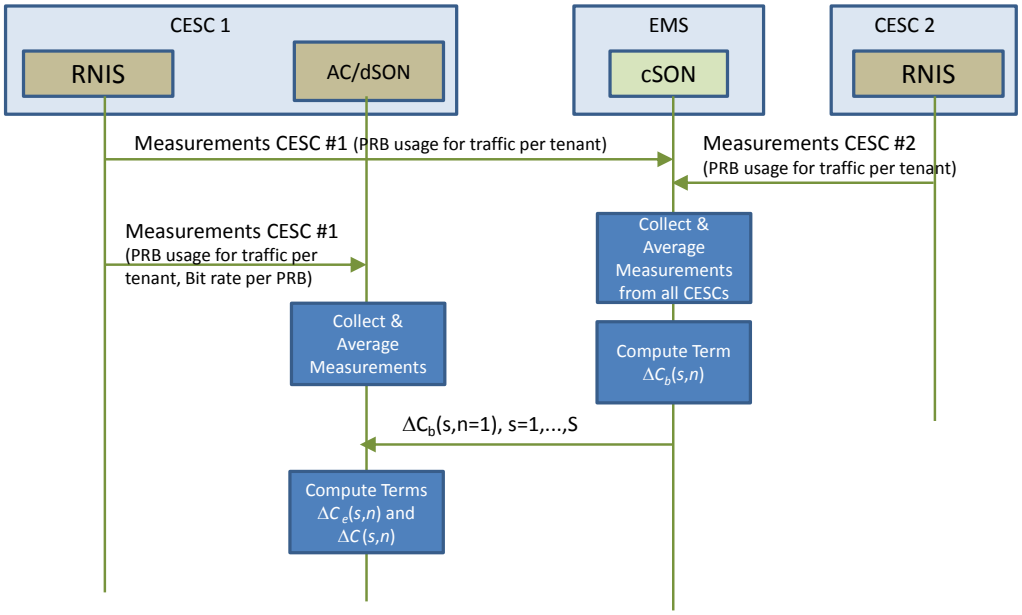

Figure 27: Workflow reflecting the interactions between the different components of the AC Self-Optimisation.

Table 4 illustrates the order of magnitude of the requirements associated to the VNFs involved in the AC Self-Optimisation process. Computations are based on the averaging window parameters defined in [29] for a scenario with 2 tenants. In particular, it is assumed that PRB usage measurements are provided by the RNIS in periods of 0.1s, while the dSON should store multiple samples and average them to makes adjustments of the thresholds every 300s. The computational requirements of the AC checks are dependent on the RAB arrival rate in a cell, for which a rate of 1 request/s has been considered for each tenant.

Table 4: Estimation of the requirements for the VNFs involved in the AC/dSON process

	Computation	**Memory**	**Networking**
AC/dSON VNF	~ 100 ops/s	~ 10 KB	~ 200 b/s (RNIS-AC)
RNIS VNF	~100 ops/s	~ 10 KB	~ 1 kb/s (SC PNF - RNIS)

3.6.4 Performance assessment

Some performance evaluation results of the proposed AC Self-Optimisation framework were presented in section 5.2 of deliverable D3.2 [1] in a scenario with 2 cells and 2 tenants. In order to further complement this evaluation, this section provides some additional results obtained in a more complex scenario with 19 cells and 4 tenants.

Evaluation is performed by means of system-level simulations in the outdoor Urban Micro scenario of [35] with hexagonal layout and considering a total of 19 cells, numbered as shown in Figure 28. All the cells operate in the same LTE carrier. The detailed simulation parameters are listed in Table 5. Simulations consider the downlink direction with S=4 tenants, denoted as s=1, 2, 3, 4. The capacity contracted by each tenant in the scenario is given by the SAGBR values in Table 5, where also the nominal capacity shares *C*(s) are presented.

The UEs of each tenant request Guaranteed Bit Rate (GBR) E-RABs following a Poisson arrival model and exponential session duration. The session generation rate of the s-th tenant is set so that its total offered load in the scenario equals SAGBR(s). However, to analyse the effect of the different spatial distributions, it is assumed that this total offered load is spatially distributed for each tenant following a Gaussian distribution centred at the colour points shown in Figure 28, which are located at distance R(m) from the centre of the scenario. The Gaussian distribution is characterized by a standard deviation of *D*(*m*) that determines how concentrated the traffic is around the central point.

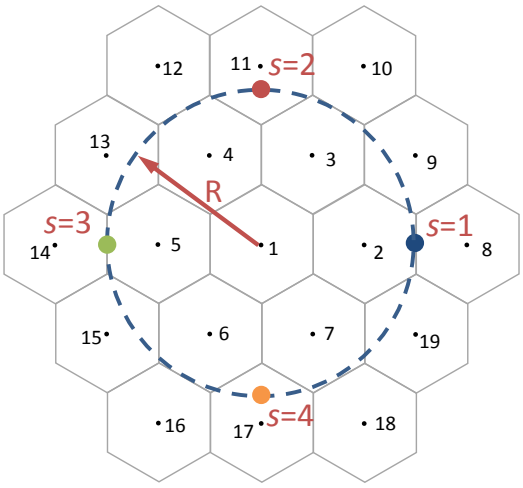

Figure 28: Considered scenario.

Table 5: Simulation parameters

Parameter	Value
ISD (Inter-Site Distance)	200m
Path loss and shadowing model	Urban micro-cell model with hexagonal layout (see details in [35])
Shadowing standard deviation	3 dB in Line Of Sight (LOS) and 4 dB in Non Line Of Sight (NLOS) [35]
Base station antenna gain	5 dB
Frequency	2.6 GHz
Transmitted power per PRB	24 dBm
Number of RBs per cell $\rho(n)$	50 RBs (1 LTE carrier of 10 MHz)

UE noise figure	9 dB
Link-level model to map Signal to Interference and Noise Ratio and bit rate	Model in section A.1 of [36] with maximum spectral efficiency 4.4 b/s/Hz.
SAGBR	$SAGBR(1)$=56Mb/s; $SAGBR(2)$=42 Mb/s; $SAGBR(3)$=21Mb/s; $SAGBR(4)$=21 Mb/s.
Nominal capacity shares $C(s)$	$C(1)$=40%; $C(2)$=30%; $C(3)$=15%; $C(4)$=15%.
GBR (R_{req})	1024 kb/s
Average session duration	30 s
$\alpha_{th}(n)$	0.75
Averaging windows	T=10s, T_e=30s, W_{cSON}=W_{dSON}=300 s
θ	0.77
Simulation duration	10000 s

The performance of the proposed hybrid SON approach is compared against a reference case where no SON technique is used, so $\Delta C(s,n)$ is set to a fixed value $\Delta C(s,n)$=0. Figure 29 depicts the total throughput improvement in % obtained by the proposed approach with respect to the reference case. In order to capture different spatial traffic distributions, results are presented as a function of the distance $R(m)$ (see Figure 28) and for different values of the standard deviation D(m). For comparison purposes, also the homogeneous spatial distribution is presented, in which the total offered load of a tenant is equally distributed in all the cells. Results reveal that the proposed approach is able to achieve substantial improvements with respect to the reference case thanks to the dynamic adaptation of the parameter $\Delta C(s,n)$ for each cell. It is also observed that the achieved gains increase when the spatial traffic distributions of the different tenants exhibit a higher degree of complementarity, i.e., for high values of R and low values of D, which correspond to situations in which the traffic of each tenant is concentrated in a different region of the scenario. In such cases, the observed gains are up to 65%.

To gain a further insight in the behaviour of the algorithm, Figure 30 presents the throughput of tenant s=1 in the different cells of the scenario when the spatial traffic distribution is given by R=300m, D=200m. It is observed that the highest benefits from the proposed SON approach are obtained in cells 2, 8, 9 and 19, which correspond to the cells surrounding the central point of the traffic distribution for this tenant (see Figure 28) and therefore they are the cells that need to serve more load from this tenant. In turn, the differences between the two approaches in the rest of the cells are smaller.

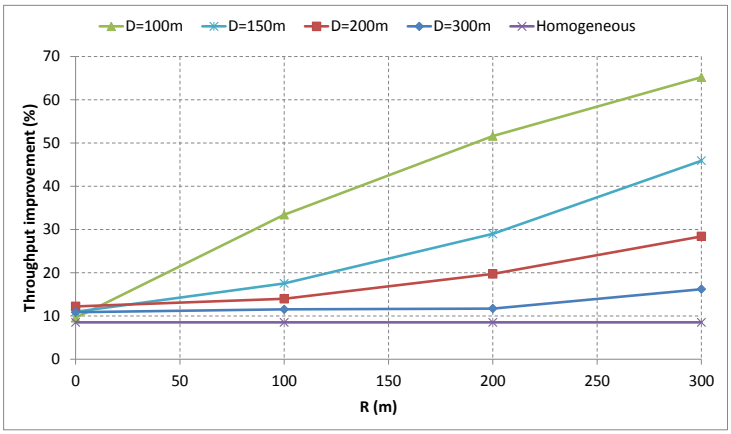

Figure 29: Throughput improvement with respect to the reference case as a function of the distance *R*(*m*) and for different values of *D*(*m*).

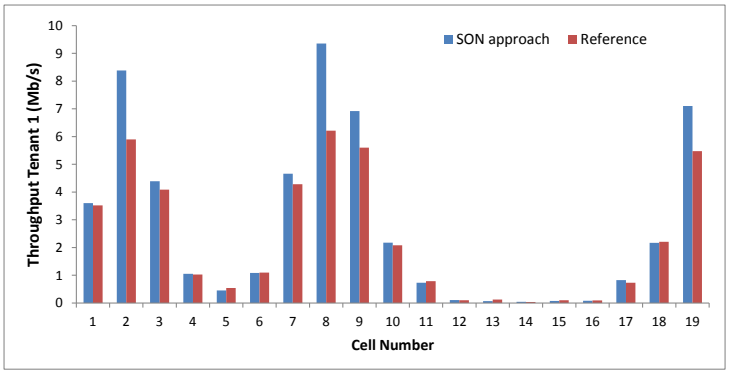

Figure 30: Throughput of tenant *s*=1 in each cell for the case *R*=300m, *D*=200m.

3.7 Self-Planning in the Wireless Backhaul

In i2CAT's wireless, SDN-*based* backhauling architecture three main centralized Self-Planning features can be identified:

- Creating the virtual network slices on top of the physical infrastructure;
- Backhauling path allocation;
- Adjustment of the airtime ratios for the VSCNOs.

The SLA signed with a tenant determines which of the CESCs will be usable by a specific tenant. For the backhaul, this information is necessary, as it determines over which of the physical wireless interfaces that form the backhaul-mesh virtual, tenant-specific interfaces will be enabled. The virtual interfaces belonging to a specific tenant form a virtual mesh that can be used to transport the traffic from the CESCs to the EPCs and *vice-versa*. On the other hand, the configuration of the physical devices, i.e., choosing the radio channel and transmission power of

the wireless transceivers used for the backhaul, is a task that should be carried out by the infrastructure provider by setting up a configuration file. This configuration file can be parsed by the SDN controller that then uses the NETCONF protocol[8] to apply the configuration to the backhauling devices. For this, an API that uses NETCONF to configure the wireless interfaces is made available, supporting NETCONF calls that are translated to calls of the platform specific tools (like *iw*) to set the transmission power, operating frequency, etc.

The task of path allocation performed by the SDN controller is to choose a (multihop) path for the data flow from an SC to a tenant's core network and to program the correspondent rules in the backhauling nodes. An algorithm that takes into account the current state of the wireless backhaul in terms of ongoing transmissions and channel loads is responsible for determining which wireless links should be used, so that data can be exchanged between the UEs and the EPC, while taking into account different performance metrics: load-balancing, maximum per path throughput, etc.

On the other hand, in order to assure the airtime ratios are assigned as determined by the SLA, the SDN controller is configured via its RESTful API. Then, the SDN controller uses the measured information about airtime usage by the registered tenants across the backhaul and readjusts the ratios applied in each of the backhaul nodes. Both Self-X features that can be categorized to be of a centralized nature are detailed in deliverable D3.2 [1]. In the following, the technical details of these to features are provided.

3.7.1 Implementation Details

The creation of the virtual network slices for each tenant happens at the very moment a tenant registers to the CESCM and the details of the SLA are handed over to the wireless backhauling system. In particular, the backhaul disposes of a Northbound API that is responsible for translating the request, i.e., the CECSs assigned to a tenant, to the NETCONF calls that then are used to launch the virtual topology on top of the physical one. The scripts performing these tasks are written in bash and the C programming language and are designed for Ubuntu-*based* environments.

The path allocation and network slicing adjustment, are implemented as part of an OpenDayLight (ODL)[9] Java bundle. The bundle is compatible with the Helium version[10] of ODL and it can be loaded dynamically during network operation. It should be noted, that this bundle is currently ported to a newer ODL version (Boron[11]). Since the bundle implements basic features of the backhauling architecture, it can be considered a core component which always needs to be active in order for the backhaul to operate correctly.

Since the controller is written in Java, it can run on any platform that supports java, thus being a very flexible solution for an SDN controller. As a major requirement, it needs to run on a machine with sufficient RAM and a powerful CPU in order to provide enough processing capacities to perform the controller tasks, which include the monitoring and maintenance of the network status, running Self-X algorithms and interacting with the nodes. The software has been tested on virtual and physical machines with at least 4 GB of RAM and Quad-Core CPUs. Since

[8] For further information see, *for example*: https://en.wikipedia.org/wiki/NETCONF
[9] For more details see: https://wiki.opendaylight.org/view/Installing_OpenDaylight
[10] For further details see: https://www.opendaylight.org/software/downloads/helium
[11] See: https://www.opendaylight.org/odlboron

the controller can run on multiple physical platforms but also in VMs, it can be implemented as both PNF and VNF. The bundle is still a work in progress: the path allocation feature is already implemented, whereas the network slicing adjustment is still in work.

3.7.2 Mapping to the SESAME components

The virtual subnetworks generated on top of the physical infrastructure are all implemented as PNFs and they run in the CESC. Further, an SDN Controller needs to be assigned to each CESC cluster in order to manage its wireless backhaul. Since the controller can be deployed as a VNF, but also as a PNF, it can be placed at multiple points in the SESAME architecture: As a VNF placed in the Light DC, in the infrastructure that hosts the CESCM, or even outside of the internal components of the SESAME architecture (CESC, CESCM), e.g., as a service in the cloud. Latter two placements are the ones that are considered for the final placement of the SDN controller. Placing the SDN controller in the Light DC would reduce the amount of computing resources available for tenant VNFs and thus the Service Function Chaining (SFC) services. Like the SDN controller responsible for the SFC, the SDN controller for the wireless should run on a dedicated server to guarantee enough resources or as a VM with sufficient resources in the cloud. Latter also results in a more flexible solution, since the VNF is not bound to a specific physical machine.

In order to test the performance achieved when running the controller remotely while being connected to a wireless backhauling infrastructure, a remote VM with a public IP was set up and connected to wireless backhaul testing facilities in different locations (i2CAT premises, external WiFi-*based* testbeds, etc.). The experiments revealed only small additional (sub-second) delays in the communication process between the SDN controller and the nodes of the backhaul. These delays do not affect the performance of the data plane, only the control plane, where sub-second delays represent no hurdle for the network's operation.

3.7.3 Workflow of the feature

As detailed in deliverable D3.2, the Self-Planning features for the wireless backhaul in SESAME has a Northbound API that allows to set up new main and backup paths between the SCs and the core networks. These backhauling "tunnels" are used to transport the data packets of a specific tenant across the backhaul. The setup of such a tunnel requires the input of two IP addresses that correspond to the two endpoints of the tunnel, i.e., the SC and the EPC of the tenant. Further, since a SC may host multiple tenants, it is also required to handover for which tenant the tunnel needs to be established. This is necessary, since multiple tenants may use the same SC (and its network IP), but will have different EPCs. Internally, the controller therefore assigns specific rules that will only apply to a specific tenant from one SC to its EPC. This is achieved by the use of tags to mark tunnels belonging to a specific match of tenant and end-to-end connection. As tag, the source IPs of the CESCs can be used, but also other alternatives like a per-GTP TEID matching are possible (even though in the context of a per-tenant routing, rather than a per-UE routing, the use of the CESC IPs is necessary).

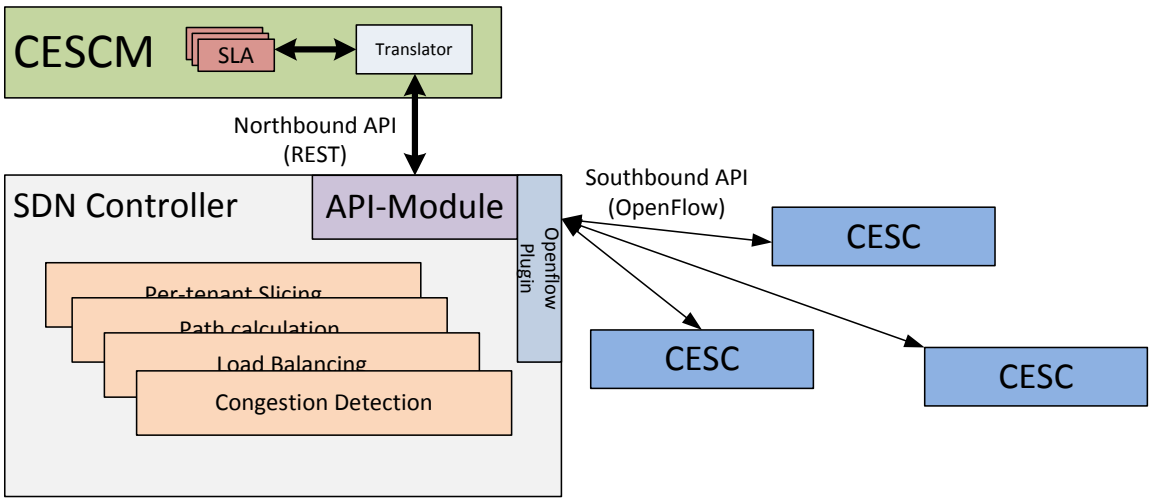

Figure 31: The SDN controller module responsible for managing the data paths of the SDN-*based* wireless backhaul is exposed to the CESCM. The SDN controller applies the configuration handed over by the CESCM to each of the CESCS in the cluster via OpenFlow.

As shown in Figure 31, the Northbound API is exposed to the CESCM. Via the API, the end-to-end (E2E) tunnels can be configured. This is done according to the requirements of the tenants that are registered at the CESCM: Every node that is used by a tenant needs to be part of the topology (i.e., connected to the mesh) and needs access to a gateway in order to reach the tenants core network. A proposal for the yet not finalized API can be found in Section 9.3 of D3.2 [1]. As detailed there, the API module is also capable of sending notifications to the CESCM about the status of the backhaul.

Since the assignment of airtime ratios as part of the virtualization of the wireless backhaul is also part of the Self-planning Self-X feature, the Northbound API offers calls to configure the ratios to be applied by each tenant. As detailed in deliverable D3.2, each tenant can be assigned a share of the wireless backhaul that is applied over the course of operation. A call to the API of the controller requires the name of the tenant and the ratio that should be applied in the backhaul. The SDN controller then takes care of configuring the SCs correspondingly via its internal NETCONF API. As part of the Self-Optimisation features, ratio violations are reacted upon by keeping track of the reported ratios during network operation and by readjusting the ratios in order to fulfil the predefined ratios. The interactions between the CESCM, the SDN controller and the CESCs for the Self-Planning and Self-Optimisation features of the airtime ratio management system are shown in Figure 32. Another (optional) configuration parameter that could be configured by the CESCM, is the radio of the wireless transceivers of the network. As mentioned before, this is currently not considered.

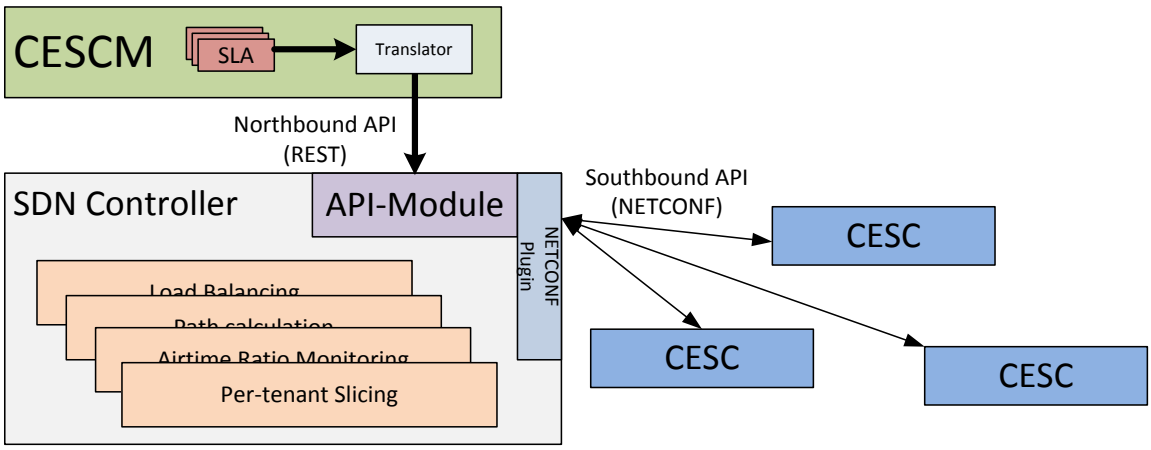

Figure 32: The SDN controller module responsible for managing the airtime ratios of the SDN-based wireless backhaul is exposed to the CESCM. The SDN controller applies the configuration handed over by the CESCM to each of the CESCS in the cluster via NETCONF.

3.8 Self-Optimisation in the Wireless Backhaul

During the operation of the wireless backhaul, several Self-Optimisation features assure the correct operation of the network. As described in deliverable D3.2 [1], the wireless backhauling architecture applies Self-Optimisation by allowing the reallocation of data flows by applying a local optimisation to the network slicing mechanism. The dynamic path allocation is a cSON feature applied by the controller. The state of the network is continuously monitored and data flows from the SCs to the EPCs for different tenants are evaluated for their impact on the overall network performance. If, according to a performance metric avoiding congested channels, a critical network situation is detected, the SDN controller tries to reallocate data flows over alternative routes in the backhaul to improve the situation. An example would be to spread different data flows over different radio channels to avoid interference between transmissions of different SCs.

On the other hand, the virtual wireless link ratio adaptation mechanism performed on a per-CESC basis is a hSON feature that tries to optimize the airtime assigned to tenants on wireless interfaces. Basically, the mechanism reassigns unused airtime to the tenants that have active transmissions to optimize the usage of the wireless resources; this is done locally at each node. On the controller side, the reports about the used airtime are collected to apply the cSON part of the mechanism by adjusting the ratios if the predefined minima for each tenants are not applied correctly at a local level.

3.8.1 Implementation Details

The data path allocation mechanism is part of a software bundle running in the ODL-*based* SDN controller. The bundle collects the periodical statistics from the wireless backhaul nodes, gathering information about the channel load measured on each wireless device, the amount of traffic sent over the devices, the amount of traffic of a tenant's data flow, etc. The information is stored on a local database, which can be consulted in order to retrieve historical information about the network. The algorithms that decide whether a critical network situation needs to be resolved check every time new statistics arrive whether any threshold for the active

performance metric (e.g., maximum channel load) is exceeded. If so, the algorithm calculates the impact of each data flow on the current network situation and tries to estimate whether a reallocation of the flow can improve the network situation. If so, new rules are installed on the virtual switches located in the SCs in order to redirect the traffic flows. Like the other Self-X features implemented for the wireless backhaul and running in the controller, they form part of a software that can either be implemented as VNF or PNF.

On the other hand, the dSON ratio adaptation algorithms applied to adjust the radio slices in the backhaul run on the SCs as a local agent software. The software is composed of a linux kernel module sitting on top of the mac80211 kernel module[12] and a NETCONF agent that allows to configure the module. These components run as PNF in the kernel and user space, *respectively*. The software has been tested on x86[13] and ARMv7[14] architectures. The cSON ratio adaptation algorithms run in the SDN controller: The reports gathered by the nodes of the backhaul allow the controller to determine whether the ratios are correctly applied on a global scale. If not, the controller uses NETCONF messages to reconfigure the ratio setting on the affected nodes.

Further details on the implementation of the Self-Optimisation features provided in the wireless backhaul are given in [1].

3.8.2 Mapping to the SESAME components

The Self-Optimisation features running in the SDN-controller are all handled internally by the software bundle and they don't require any interaction with the rest of the architecture. Since it is part of the ODL SDN-controller, this cSON feature will run wherever the controller is placed, e.g., as a VNF within the SESAME architecture or as a cloud service running on a dedicated server. On the other hand, the Self-Optimisation applied to the virtualization of the backhaul are part of the CESC running on the same hardware as other PNFs, such as the SC-PNF.

3.8.3 Workflow of the feature

Since the Self-Optimisation features for the wireless backhaul do not generate any output or require any input during runtime, the interactions with the rest of the architecture are limited. At the current state of development, there are several parameters like the congestion threshold limit that are static and cannot be changed during runtime. These values are read from a configuration file located at the SDN-controller that is parsed during the initialization of the software bundle. These are the main parameters that can be configured in the configuration file:

Table 6: Configuration parameters for the wireless backhaul controller

Parameter	Description	Type
Channel Load Threshold	This threshold determines upon which value the SDN	Double

[12] For further details see: *https://wireless.wiki.kernel.org/en/developers/documentation/mac80211*

[13] x86 is a family of backward compatible instruction set architectures based on the Intel 8086 CPU and its Intel 8088 variant. More relevant information can be found, *inter-alia*, at: *https://en.wikipedia.org/wiki/X86*

[14] For further information see, *for example*: *https://en.wikipedia.org/wiki/ARM_architecture*

	controller tries to apply rerouting of the traffic flows in order to balance the network load (60% by default)	
Wait Cycles	Once a flow has been redirected, it may take some time for the changes to be monitored by the controller. In order to avoid the controller from immediately reconfiguring the network again after a change has been made, it can be forced to wait X statistic cycles before allowing new rerouting of traffic.	Short
Historical Weight (BETA)	Every time a new statistic is received, a moving average is calculated for the observed channel load, Tx rates, etc. The factor BETA determines the impact of the new values on the moving average (0.25 by default)	Double

3.9 Self-Healing in the Wireless Backhaul

The Self-Optimisation mechanisms running in the SDN-controller and in the SCs are responsible for maintaining (a high) performance of the wireless backhaul during operation. Still, it is possible that due to hardware issues or due to other external effects, such as radio interference, parts of the backhaul become unusable. In order to cope with situations where entire backhaul nodes or wireless links of the backhaul fail, a set of cSON and dSON Self-Healing mechanisms are necessary.

The SDN controller monitors the state of the wireless backhaul and therefore has knowledge of the aforementioned issues. Link breaks or node failures result in the affected entities disappearing from the global view of the controller, as they no longer form part of the network. Whenever a link or a node disappears, the controller is responsible for checking if any traffic flows are affected by the failure. If so, these flows are redirected over an alternative, valid route.

In order to improve the reaction time to such network failures, a local, dSON mechanism has been designed that is capable of detecting link issues a wireless transceiver may have. If such an issue is detected, e.g., due to consecutively missing beacons from another peer or when exceeding a certain amount of retransmissions, the local recovery mechanism redirects data flows over the backup paths that have been pre-provisioned by the controller as part of the Self-Planning mechanism.

Further details on the Self-Healing features are provided in Section 9.2.3 of deliverable D3.2.

3.9.1 Implementation Details

Since the SDN-controller is aware of the backhaul topology at any point of time, it can detect when a link or a node disappears from the topology. The functions that analyse whether a data flow is affected by the network issues and then recalculates paths if necessary is part of the software bundle that runs in the SDN-controller (like the other Self-X mechanisms of the backhaul that run in the controller). Whenever a change in the topology is detected, the software bundle runs the methods to check whether the data plane of the backhaul is affected. The methods called to reassign data flows are the same that are used to allocate new end-to-

end flows in the backhaul. The software bundle has been tested as PNF and as VNF running on x86 architectures, but it also capable of running on other hardware that supports java and has enough computer resources.

The Self-Healing mechanism responsible for detecting link failures locally is implemented as an agent software that is composed of two main parts: bash and C scripts that keep track of the state of the virtual switch and the mac80211 module with some modifications that report changes in the state of the wireless links. This mechanism has been developed and tested on various x86 and ARM platforms (v6/v7) as PNF and is an optional feature that can be enabled or disabled at any point in time.

3.9.2 Mapping to the SESAME components

The cSON feature runs in the SDN-controller that can be placed in several points of the SESAME architecture that has access to the backhauling network: the Light DC or the CESCM. Again, it is not recommendable to run it in the Light DC, since its resources should be dedicated to the VNFs for the tenants. Instead, it should be running as part of the CESCM or even a dedicated server, locally or in the cloud.

The dSON Self-Healing feature needs to run on the SC, where it can monitor the state of the wireless links by checking the status of the Wi-Fi interfaces and where the agent software will be running.

3.9.3 Workflow of the feature

The input necessary to decide whether there is a network failure in the backhaul is coming from the virtual network switches (i.e., the SCs): the SDN-controller polls the wireless backhaul statistics from each of the switches. If a SC is no longer reachable due to a hardware issue or a link failure, the reports will either no longer arrive at the SDN-controller or they will report about the change of topology. No output for other SESAME components is generated, since the Self-Healing mechanism resolves the issues internally. Similarly, the local Self-Healing applied at the SCs does not require or generate any output for other SESAME components.

3.10 Self-Planning in Multi-Tenant Small Cell Networks

A new framework for Self-Planning considering the specificities of multi-tenant small cell networks was presented in Section 4 of Deliverable D3.2 [1]. It focuses on the automation of the network planning tasks for identifying the parameter settings of new network elements, including the site locations and the channel assignment to the different cells. The proposed framework includes a functional architecture for introducing multi-tenancy in the Self-Planning process, together with a specific algorithmic solution for deciding the required updates in the network layout and the channel assignment to the small cells.

The Self-Planning function is a centralized SON functionality, because it needs to make decisions based on the overall set of small cells deployed in a certain area by a SCNO. Besides, the developed Self-Planning function addresses the use case “Optimized Radio Network Capacity Planning and Operation mechanisms of the Small Cell Network Operator (SCNO)” defined in deliverable D2.1 [28].

3.10.1 Implementation Details

The functional architecture for the considered Self-Planning framework is depicted in Figure 33 and includes the following entities (see Section 4 in [1] and [37] for further details):

- Multi-tenancy management entity: It acts as an interface between the tenants, i.e., the VSCNOs, and the network planning activities of the network provider, i.e., the Small Cell Network Operator (SCNO). Essentially, this entity translates the Service Level Agreement (SLA), which is normally expressed in terms of aggregate (or average) values over relatively coarse time and space scales, in terms of detailed planning specifications that reflect the capacity contracted by the tenant in each small cell and time.
- Self-planning entity: This entity checks whether or not the currently deployed capacity in the small cell network fits with the actual tenants' demand and identifies the required changes in the network layout and channel allocation. The first check is performed by the capacity conformance monitoring module, which compares the current deployed bandwidth in each small cell with the actual required bandwidth. If the ratio between the two is above a certain threshold, then the capacity dimensioning and planning module is triggered. This module executes an iterative algorithm to determine the optimal network configuration to cope with the existing traffic demand. For that purpose, it may suggest the addition or removal of channels to the different small cells, the addition of new small cells or the removal of existing small cells. This module relies on the network performance model to estimate the performance of each candidate solution considered by the iterative algorithm without the need of implementing the solution in the real network.

As shown in Figure 34, both the multi-tenancy management entity and the Self-Planning entity can be supported by a traffic forecasting entity, which provides the predicted traffic demand at a future time in every small cell. This enables a proactive execution of the Self-Planning process in order to anticipate the requirement of future changes in the network. This anticipation becomes "key" as long as the deployment of new infrastructure may require substantial time compared to the evolution of the traffic demand.

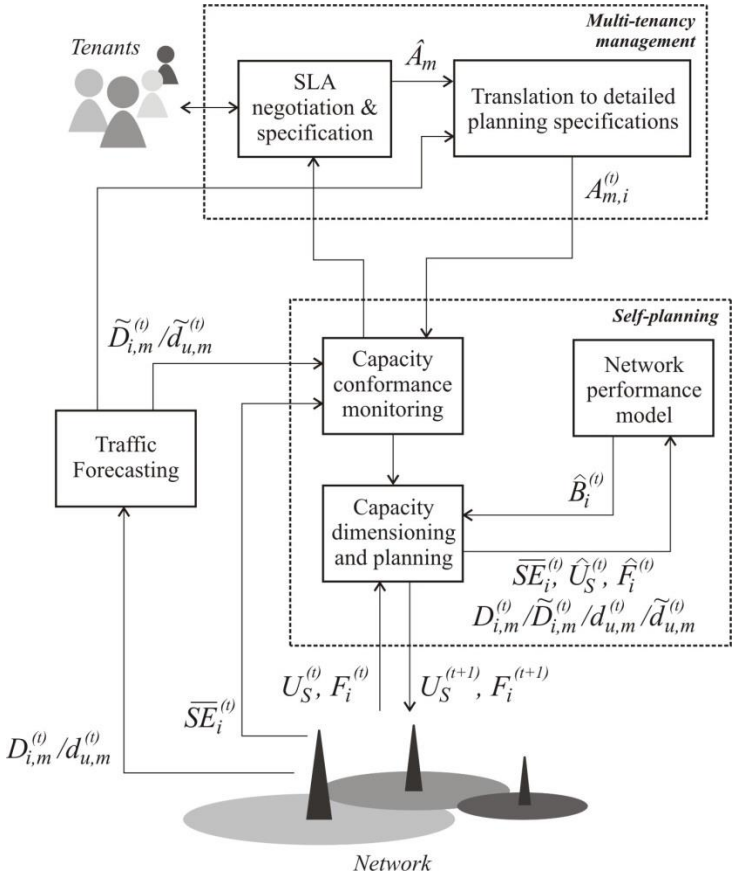

Figure 33: Functional architecture of the Self-Planning framework

3.10.2 Mapping to the SESAME components

Being a cSON function, the abovementioned entities of the Self-Planning process run at the management systems. Then, in the context of the SESAME architecture of [30], it is envisaged that both multi-tenancy management entity and Self-Planning entity are associated either to the Element Management System (EMS) of the CESCM or to the Network Management System (NMS).

3.10.3 Workflow of the feature

According to the detailed formulation and system model presented in section 4 of [1], as well as the architecture shown in Figure 33 the proposed Self-Planning framework requires the following input measurements:

- Actual traffic demand of each tenant m in each small cell *i* ($D_{i,m}^{(t)}$), measured in Mb/s: This measurement can be computed at the EMS from the PM reports provided by each CESC. Specifically the traffic demand is estimated as the served traffic divided by the accessibility probability (i.e., the probability success rate at E-RAB establishment). For that purpose, the served traffic can be obtained from the IP throughput measurements of [32]. In turn, the accessibility probability is defined in [38] based on the E-RAB and RRC connection related measurements of [32].

- Estimation of the average spectral efficiency in each small cell *i* ($\overline{SE_i^{(t)}}$): This is estimated as the ratio between the total transmitted bit rate in the *i*-th small cell and the amount of Physical PRBs that have been used for transmission. The bit rate is obtained from the IP throughput measurements of [32] while the number of used PRBs is obtained from the total PRB usage measurement of [32]. Both measurements are included in the PM reports sent by each CESC to the EMS.

The abovementioned measurements are to be computed as average values in the long-term. Therefore, the required periodicity for providing the information to the EMS will be long, e.g., typically in the order of 10-15 min.

In addition to the above measurements, the Self-Planning entity also requires to have knowledge about the current network configuration (e.g., locations of the deployed small cells $U_S^{(t)}$ and channels allocated to each small cell $F_i^{(t)}$). This information can be extracted from the configuration of each small cell, stored in the CESC managed objects of the EMS (see [39]). Similarly, the information about the negotiated SLA with each tenant *m*($\hat{A}_m$) can also be obtained from the SLA managed object defined in [39], also stored at the EMS.

Figure 34 illustrates a workflow that shows the interactions between the different components of the proposed Self-Planning framework. They are described in the following:

1) The multi-tenancy management entity translates the overall SLA of a tenant *m*($\hat{A}_m$) defined in terms of the overall capacity to be provided across all the cells into detailed planning specifications for each cell *i*, ($\hat{A}_{m,i}^{(t)}$). This process is executed in the long-term only when there are changes in the agreements with the involved tenants (e.g., due to the addition of a new tenant, the modification of the SLA, etc.). The result of the translation is delivered to the Self-Planning entity.

2) The Self-Planning entity collects the PM reports of the different CESCs in the scenario. The PM reports can be generated by the SC-PNF and/or by the SC-VNF depending on the considered functional split. They include the measurements listed above from which the Self-Planning entity can compute the relevant metrics (traffic demand per tenant and average spectral efficiency in each cell).

3) The Self-Planning entity triggers the capacity conformance monitoring check according to eq. (5) of [1] (see [1], [37] for details). If the check is passed, no modification is made in the network and the process will be repeated again in the future according to some specific long term periodicity (e.g., in the order of some hours). Instead, if the check is not passed, the capacity dimensioning and planning module is triggered. It executes the algorithm presented in section 4.2 of [1] and in [37] to identify appropriate modifications to be made in the network (addition/removal of small cells, changes in the assigned channels to each small cell). The outcome will be a recommendation for the SCNO that will be in charge of implementing the change. This implementation can be automated, e.g., in case of a change in the channel allocation, or it can be manual, e.g., when a new small cell has to be deployed.

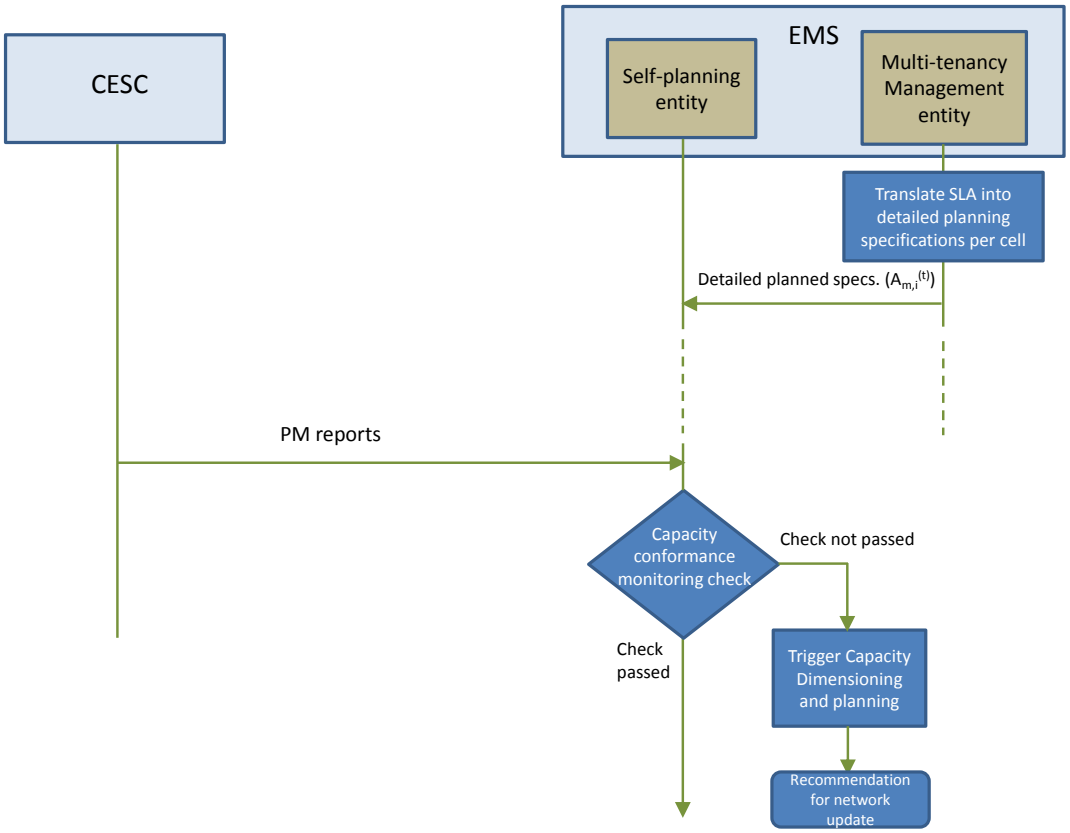

Figure 34: Workflow reflecting the interactions between the different components of the Self-Planning framework.

3.10.4 Performance assessment

This section extends the results on the Self-Planning framework that were presented in section 4.3 of [1] with additional new results considering different traffic situations. The results consider the planning of the network whenever a new tenant is added.

3.10.4.1 Simulated scenario

An urban SC scenario with dimensions 0.4 km × 0.4 km and a grid resolution of 5 m has been considered. To represent the areas where deploying SCs is possible, e.g., no backhaul and site acquisition constraints, 2% of the points (or pixels) in the scenario have been randomly selected as candidate site locations. The actual network layout and the traffic demand at the busy hour in the situation before the consideration of the new tenant are represented in Figure 35, where the triangles represent the location of the 3 deployed Small Cells (SCs) and the values in brackets are the number of allocated channels. Colour scale indicates the traffic demand density, which is non-uniformly distributed over the considered area. The traffic demands supported by SCs 1-3 are 7.3, 4.6 and 4.1 Mbps, respectively.

The path loss is computed using the International Telecommunication Union (ITU) Indoor Hotspot (InH) model in [35]. The carrier frequency is 5 GHz. The frequency band is composed of 4 channels, with channel bandwidth B=20 MHz. The SCs are assumed to be omnidirectional with an antenna gain of 2 dB. The transmit power P_i, which can vary between 24 and 10 dBm, is configured for each SC to have SINR equal to 9 dB for a UE located at $\sqrt{3}/2$ of the Inter-Site Distance (ISD) [40]. The terminal noise figure is 9 dB. The spectral efficiency function *SE(SINR)*

used to compute the average spectral efficiency $\overline{SE_i^{(t)}}$ at SC *i* depending on the SINR at each pixel is obtained from Section A.1 in [36] with maximum spectral efficiency of 4.4 b/s/Hz.

From a network planning perspective, the parameters used in the *capacity conformance monitoring* module to trigger the planning actions are configured as α=0.9, β=0.7 and γ=0.05 (see [1][37] for details). In addition, the maximum number of allocated channels per SC is set to K_{max}=2, while the maximum number of SCs that can be deployed in the considered area is set to N_{max}^{SC} = 8.

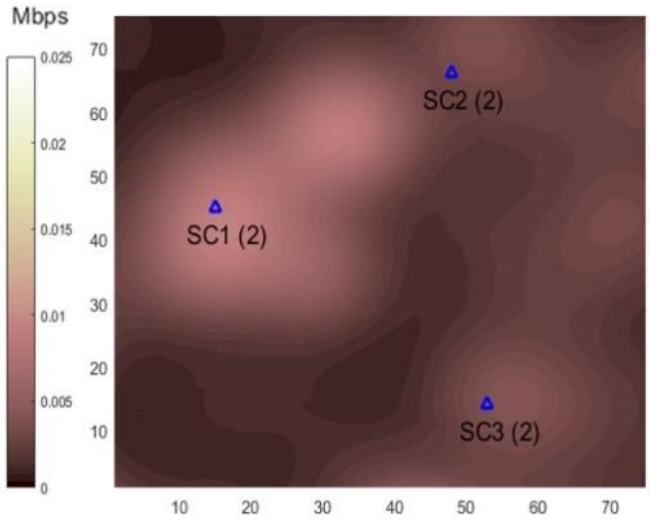

Figure 35: Traffic demand and network deployment in the initial situation (before new tenant's arrival).

3.10.4.2 Analysis of the network planning solutions

Let assume that the SLA of the new tenant *m* is translated to a specification at the busy hour of $A_m^{(t_B)}$=20 Mbps. At this initial stage, the new tenant's spatial traffic demand distribution is assumed to be unknown, so the planning is carried out using the detailed planning specifications from the following methods (see [1], [37] for details):

- Uniform distribution: In case that the spatial traffic demand of the new tenant is unknown, an even distribution among the SCs is assumed. Estimation can be conducted in two different ways:
  - At SC-level: In this case the total traffic demand $A_m^{(t_B)}$ is uniformly distributed between all the deployed small cells.
  - At pixel-level: In this case the total traffic demand $A_m^{(t_B)}$ is uniformly distributed between all the pixels existing in the area.
- Correlated distribution: In case that correlation between the traffic demand for the new tenant and the already existing tenants is expected, areas with higher traffic demand of other tenants will receive a greater contribution of $A_m^{(t_B)}$. In this way, the estimation of the traffic demand of the new tenant can be based on the traffic demand of the other tenants. Again, two different options are possible:
  - At SC-level: In this case the total traffic demand of the new tenant $A_m^{(t_B)}$ is distributed among the deployed SCs following the same distribution of the traffic demand of the other tenants.

- At pixel-level: In this case, traffic measurements at pixel-level are taken from call traces that provide geo-location information for each user of the existing tenants in an automatic way. This allows knowing the traffic distribution of the existing tenants in each pixel of the scenario. Based on this, the total traffic demand of the new tenant $A_m^{(t_B)}$ is distributed among the different pixels following the same distribution of the other tenants.

Thus, the total traffic demand is calculated as the actual traffic demand from existing tenants plus the estimated new tenant's demand. After generating the detailed planning specifications, it is observed that in the *capacity conformance monitoring* module condition is satisfied for the 3 deployed SCs, so that the *capacity dimensioning and planning* module is launched.

Figure 36(a-d) shows the results of the planning process for different sets of detailed planning specifications corresponding to the four abovementioned planning methods. The number in parenthesis close to each SC indicates the number of allocated channels.

For the uniform distribution at the SC-level method, Figure 36 (a) shows that, given that SC 3 has the largest service area, the traffic demand per pixel in this SC is estimated to be lower than in SC 1 and 2. Then, the capacity dimensioning and planning module adds 3 new SCs, which are located close to the border of the service areas of the existing SCs, so that the spectral efficiency is mainly improved at the cell-edge.

The correlated distribution at the SC-level method is represented in Figure 36 (b). This method estimates that SC 1, which initially carried more traffic (i.e., 7.3 Mbps), is the cell that receives proportionally more traffic from the new tenant than other SCs. Therefore, compared to the previous method, it is observed that SC 4 is located further inside the service area of SC 1, whose traffic demand is higher than in the first method.

For the method based on uniform distribution at the pixel-level, illustrated in Figure 36 (c), it locates SC 6 in an area with high traffic density thanks to the higher resolution in the spatial distribution of the traffic demand. However, the assumption of uniform traffic distribution for the new tenant has raised the number of allocated channels in the scenario to 12.

The last method (see Figure 36 (d)), based on correlated distribution at the pixel-level, produces the largest variations in the traffic demand per pixel. As a result of the planning, SC 5 and 6 are placed in, or close to, the area with high traffic density, so that part of the traffic is offloaded from SC 1, having this cell only one channel allocated after the planning.

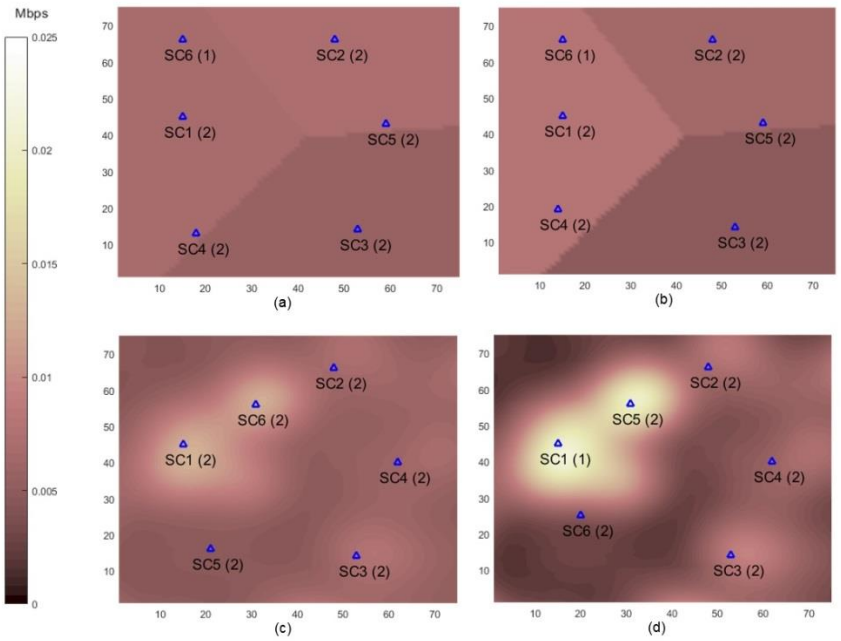

Figure 36: Network deployment and estimated traffic demand using the detailed planning specifications: (a) Based on uniform distribution at the SC-level, (b) Based on correlated distribution at the SC-level, (c) Based on uniform distribution at the pixel-level, and (d) Based on correlated distribution at the pixel-level.

As a result of each of the four planning methods of Figure 36, Table 7 shows the estimated required bandwidth in MHz in each SC based on the estimated traffic demand of the new tenant. The table shows that the method based on correlated distribution at the SC-level requires a higher bandwidth in SCs 4 and 6 compared to the method based on uniform distribution. This is due to the high traffic density in the left side of the scenario results. Regarding the uniform distribution at the pixel-level method, it is noted that the increased required bandwidth in SC 6 is due to a higher traffic demand estimated in its service area (see Figure 36 (c)). In addition, such a location for SC 6 allows offloading some traffic from SC 2, reducing its required bandwidth. The method based on correlated distribution at the pixel-level provides the lowest total required bandwidth since the traffic is estimated to be more concentrated in specific areas that are covered by nearby cells.

Table 7: Estimated required bandwidth (MHz) for each SC

SC	Uniform SC-level	Correlated SC-level	Uniform Pixel-level	Correlated Pixel-level
1	19.0	17.8	18.9	15.7
2	21.0	19.3	15.1	13.6
3	22.6	22.1	19.3	17.2
4	18.4	21.1	15.0	12.3
5	22.6	21.0	20.2	13.1
6	11.6	12.3	17.3	19.5
TOTAL	115.2	113.6	105.8	91.4

3.10.4.3 Analysis of the network operation with the new tenant

This section evaluates the solutions of the planning algorithm when the new tenant's service is operative and the actual traffic demand at the busy hour of the new tenant is spatially distributed as illustrated in Figure 37, where two cases are distinguished. In the former (Figure 37 (a)), the new tenant's spatial traffic demand exhibits quite high correlation with already existing tenants, whose spatial traffic distribution is represented in Figure 35. Specifically, using Pearson's coefficient, both traffic distributions are 75% correlated. In the latter case (Figure 37 (b)), the distributions are only 15% correlated.

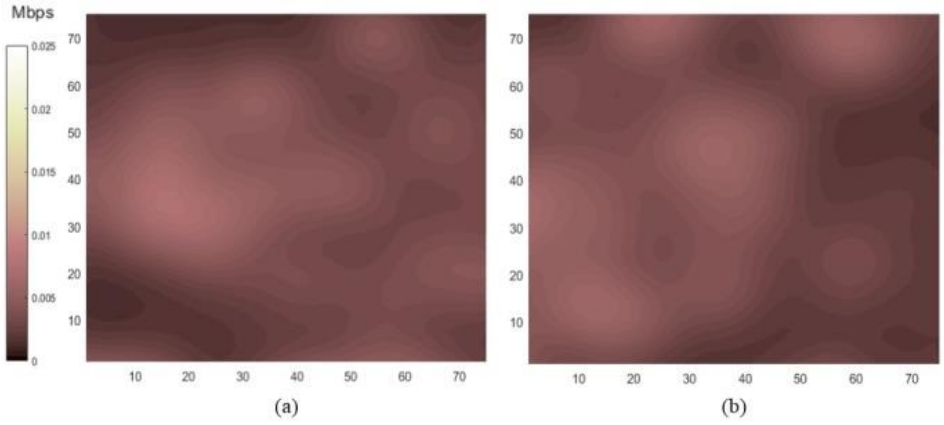

Figure 37: Traffic demand of the new tenant: (a) 75% correlated with network's traffic demand; (b) 15% correlated with network's traffic demand.

Let assume now that the network has been deployed as dictated by the planning (i.e., with the real network layouts are as illustrated in Figure 36 (a-d)) and let consider the real traffic demand of the new tenant shown in Figure 37. In that case, Table 8 shows the required bandwidth in each SC considering the actual traffic demand for the two levels of correlation with the different planning methods. The last row in the table shows values aggregated over all the SCs.

Table 8: Actual required bandwidth (MHz) for each SC for the 75% and 15% correlated traffic

Correlation (%)	SC	Uniform SC-level	Correlated SC-level	Uniform Pixel-level	Correlated Pixel-level
75	**1**	27.6	22.7	18.5	17.2
	2	18.4	18.5	12.7	12.7
	3	22.0	24.6	18.6	20.4
	4	15.3	17.1	15.4	15.2
	5	21.2	21.2	18.6	13.8
	6	9.9	9.9	15.2	24.1
	TOTAL	114.4	114.0	99.0	103.4
15	**1**	27.1	22.2	19.7	18.8
	2	20.9	20.9	14.9	14.9
	3	19.1	21.8	16.2	17.7

	4	17.4	19.3	12.5	12.3
	5	18.9	18.9	22.4	18.6
	6	11.6	11.6	20.1	32.3
	TOTAL	115.0	114.7	105.8	114.6

Note that the differences between Table 7 and Table 8 reflect the effects of the traffic estimation errors made by each planning method when the traffic of the new tenant was unknown. In this respect, the method based on correlated distribution at the pixel-level provides the largest variation in total required bandwidth between Table 7 and Table 8. This is because the new tenant's traffic demand is not as concentrated as expected in this method. Going into detail, it is observed that SC 1 experiences a high variation between the estimated and actual required bandwidth (i.e., between Table 7 and Table 8) for the methods with resolution at the SC-level. The reason for this is that the estimated traffic in this region is covered by SC 1, 4 and 6 (see Figure 36 (a-b)), while in fact the real traffic is very concentrated in the vicinity of SC 1. Another significant variation in the required bandwidth between Table 7 and Table 8 lies in SC 6 for the method based on correlated distribution at the pixel-level. Such a variation is especially high in the case of 15% correlated traffic due to the poor match between the location of SC 6 in Figure 36 (d) and the distribution of high-traffic areas in the left lower corner of Figure 36 (b).

The method that fits better the traffic demand minimizes the required resources in terms of number of SCs and/or allocated channels. According to Table 8, this does not necessarily mean a lower value of total required bandwidth. For example, the method based on uniform distribution at the pixel-level provides the lowest value, however, it is because this method allocates a higher number of channels in the scenario, as reflected in Figure 36 (c). Following this principle, in the case of 75% correlated traffic, the best method is the correlated distribution at the pixel-level, since the obtained value (i.e., 103.4 MHz) is the lowest in comparison with the methods that allocate the same number of channels.

In the case of 15% correlated traffic, determining the best method requires a deeper analysis. At a glance, the two SC-level methods and the correlated distribution at the pixel-level method provide a similar required bandwidth (around 115 MHz), meaning that these three methods achieve similar performance. However, to complete the analysis in the time domain, a new iteration of the Self-Planning process is needed after the network has been deployed with each of the planning methods. In particular, the capacity conformance monitoring check is evaluated again to determine whether there exists a lack of capacity or not. If so, the capacity dimensioning and planning module is relaunched to provide a new network configuration. The results before and after this process are presented in Table 9 for the different methods. Looking at the method based on correlated distribution at the pixel-level, it is noted that the actual required bandwidth in SC 1 is so high that the capacity conformance monitoring check would not be fulfilled. As a result of the re-planning process, a new SC is deployed, as seen in Table 9.

Table 9: Network deployment before and after re-planning (for 15% of traffic correlation)

Method	**#SCs**		**#channels**		**Req. BW**	
	Before	*After*	*Before*	*After*	*Before*	*After*
UNIFORM SC-LEVEL	6	6	11	11	115.0	115.0

CORR. SC-LEVEL	6	6	11	11	114.7	114.7
UNIFORM PX-LEVEL	6	6	12	12	105.8	105.8
CORR. PX-LEVEL	6	7	11	12	114.6	101.8

Finally, as a reference for comparison with the planning methods discussed in Figure 36, Figure 38 shows the network planning that would have been obtained if the real traffic of the new tenant shown in Figure 37 would have been known during the planning process.

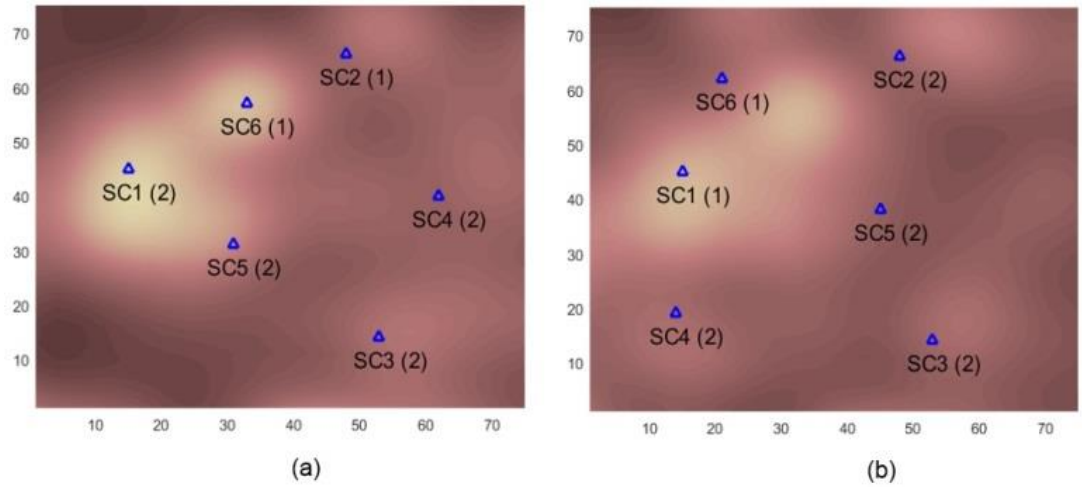

Figure 38: Network deployment if the real traffic demand of the new tenant had been known by the planning process for: (a) 75% correlated traffic; (b) 15% correlated traffic.

Regarding the case of 75% correlated traffic, it is observed in Figure 38 (a) that, in line with the methods based on correlated distribution (see Figure 36 (c) and (d)), a SC (in this case, SC6) has been placed in one of the areas with high traffic density. This differs from the methods based on uniform distribution, which are not able to correctly place a SC in such an area of high traffic density [see Figure 36 (a) and (b)]. In addition, the network layout in the reference case of Figure 38 (a) is very similar to that based on correlated distribution at the pixel-level of Figure 36 (d), where the main difference lies in the locations of SC 5 in Figure 38 (a) and SC 6 in Figure 36 (d).

With respect to the case of 15% correlated traffic [see Figure 38 (b)], it is noted that SC 4, 5 and 6 are located in areas where the traffic demand of the new tenant is higher, according to the spatial distribution shown in Figure 37 (b). Since the correlation between the actual and estimated tenant's traffic demand is low, the differences in the network layout between the reference and the proposed methods of Figure 36 are considerable, especially if the spatial resolution is provided at the pixel-level. Note also that the reference approach provides the lowest number of allocated channels (10 in total) for both correlation levels in the scenario, which is expected because it considers the real traffic distribution.

3.10.4.4 Analysis of the impact of the tenant's traffic volume

The previous analysis has been carried out for a tenant's traffic volume at the busy hour of 20 Mbps. In this section, a higher traffic volume is considered to evaluate the response of the planning methods when a considerable number of SCs has to be deployed in one execution of

the planning process. In particular, the planning specification at the busy hour has been set to 80 Mbps. In addition, the maximum number of SCs that can be deployed has also been changed to 12.

Following the same structure of Table 9, Table 10 shows the obtained network configuration in terms of number of SCs and allocated channels, as well as the actual required bandwidth, for each method when it is applied before network operation and after network operation with the new tenant (i.e., the re-planning phase). In this way, the only method that produces new variations in the updated network configuration is the correlated distribution at the pixel-level. In particular, for a 75% correlated traffic, new channels are added, while for a 15% correlated traffic, a new SC is deployed, which is a more drastic change due to the worse traffic estimation.

With high tenant's traffic volume, it is observed that the number of SCs becomes a key variable to determine the superior performance of the reference approach and the bad performance of the method based on uniform distribution at the pixel-level, since they require a different number of SCs than the other methods. The reference scheme requires the lowest number of SCs (i.e., it minimizes deployment costs). In the case of 75% correlated traffic, the best methods (without considering the reference) are those based on correlated distribution, since they require a lower number of SCs and bandwidth. In the case of 15% correlated traffic, the best results are obtained by the methods with spatial resolution at the SC-level, since the low correlation of the traffic overrides the advantages of using a higher resolution.

Table 10: Network deployment before and after re-planning

Correlation (%)	**Method**	**#SCs**		**#channels**		**Req. BW**	
		Before	*After*	*Before*	*After*	*Before*	*After*
75	***UNIFORM SC-LEVEL***	10	10	16	16	172.2	172.2
	CORR. SC-LEVEL	10	10	16	16	163.9	163.9
	UNIFORM PX-LEVEL	11	11	16	16	162.9	162.9
	CORR. PX-LEVEL	10	10	13	16	181.7	166.6
15	***UNIFORM SC-LEVEL***	10	10	16	16	173.8	173.8
	CORR. SC-LEVEL	10	10	16	16	163.3	163.3
	UNIFORM PX-LEVEL	11	11	16	16	158.6	158.6
	CORR. PX-LEVEL	10	11	13	16	220.9	170.1

4 Conclusions

This deliverable presented SESAME's framework of a distributed network management system capable to host and run Self-X functionalities, extending the work performed in the previous deliverable D3.2 that presented a selection of the most important Self-X features for the SESAME system. The framework is a key component of the SESAME architecture, as it plays an important role in the configuration and the active operation of the architecture. Self-Planning features are responsible for an organized and robust set up of the RAN, the SC PNF, a variety of VNFs of the system, and the wireless backhaul. During operation, after the set-up phase, the Self-Optimisation and Self-Healing features keep track of the systems status, taking action whenever there are issues or when there is a change to improve the performance of the system by reconfiguring components or reassigning system resources. The conclusions of the two main sections are given below.

In Section 2 it is shown how the different Self-X features form an architecture-wide system of distributed, centralised and hybrid Self-Organizing Network functionalities. An overview of the Self-X framework integrated in SESAME is given and the main task of the framework to plan, maintain and to optimize the performance of the SESAME architecture in an autonomous way is explained. A distinction is made between some Self-X features that act independently on a local level, i.e., a single SESAME component like the SC PNF, and some that interact with other SESAME components and Self-X functionalities to fulfil their tasks. For the latter case, the internal information exchange system based on standardized protocols and operations to request and store information in different system components used by the Self-X framework is presented. A network slicing architectural framework for 5G systems is presented and the fundamental design challenges for the realization of the RAN slicing are identified, revealing further details on several entry points of the framework for external actuators, such as VSCNOs, that allow the configuration of specific aspects of the RAN slicing and the SC PNF.

After giving this high-level view of the Self-X framework, a detailed, more technical description of each of the framework's components is provided in Section 3. First, for every feature technical details on the implementation and the operation are given. Since every Self-X feature is part of the SESAME architecture, each feature's mapping to the SESAME components is explained. In order to understand how each feature interacts with the system and other Self-X features, their workflows are described. The description of the workflows also reveals whether APIs are used to exchange information or if the Self-X features operate relying on other component's storage/information system. Above that, for some features additional evaluations that extend the analysis that were presented in previous deliverable D3.2 are performed.

References

[1] J. Pérez-Romero (editor), "Self-X features and virtualised CESC multi-tenancy techniques evaluation", Deliverable D3.2 of 5G-PPP SESAME project, March, 2017.

[2] ETSI GS NFV-MAN 001 (V1.1.1) Network Function Virtualisation (NFV); Management and Orchestration"" , December, 2014

[3] "ETSI Multi-access Edge Computing", available at: *http://www.etsi.org/technologies-clusters/technologies/multi-access-edge-computing*

[4] F. De Pellegrini, A. Massaro, L. Goratti and R. El-Azouzi, "A Pricing Scheme for Content Caching in 5G Mobile Edge Clouds", in Proceedings of the International Conference WINCOM, 26-29 Oct. 2016.

[5] F. De Pellegrini, A. Massaro, L. Goratti and R. El-Azouzi, "Bounded Generalized Kelly mechanism for Multi-Tenant Caching in Mobile Edge Clouds", in Proceedings of the International Conference NETGCOOP, Network Games, Control, and Optimization, 23-25 Nov. 2016.

[6] NGMN Alliance, "5G White Paper", February 2015.

[7] 5G Americas, "Network Slicing for 5G Networks and Services", November 2016.

[8] 3GPP TR 22.864: "Feasibility Study on New Services and Markets Technology Enablers - Network Operation; Stage 1 (Release 15)"", September 2016.

[9] 3GPP TS 22.261 v15.0.0, "Service requirements for the 5G system; Stage 1 (Release 15)", March 2017.

[10] 3GPP TR 23.799 v14.0.0, "Study on Architecture for Next Generation System (Release 14)", December 2016.

[11] P. Rost (editor), "Functional Network Architecture and Security Requirements", Deliverable D3.1 of the 5G NORMA project, December 2015.

[12] METIS II White Paper, "Preliminary Views and Initial Considerations on 5G RAN Architecture and Functional Design", March 2016.

[13] O. Sallent; J. Perez-Romero; R. Ferrús; R. Agusti, "On Radio Access Network Slicing from a Radio Resource Management Perspective" to appear in IEEE Wireless Communications, April 2017.

[14] 3GPP TR 38.801 v14.0.0, "Study on new radio access technology: Radio access architecture and interfaces (Release 14)", March 2017.

[15] V. Del Piccolo, A. Amamou, K. Haddadou and G. Pujolle, "A Survey of Network Isolation Solutions for Multi-Tenant Data Centers", in IEEE Communications Surveys & Tutorials, vol. 18, no. 4, pp. 2787-2821, Fourthquarter 2016.

[16] 3GPP TR 38.912 v14.0.0, "Study on New Radio (NR) access technology (Release 14)", March 2017.

[17] A. A. Zaidi et *al.*, "Waveform and Numerology to Support 5G Services and Requirements", in IEEE Communications Magazine, vol. 54, no. 11, pp. 90-98, November 2016.

[18] K.I. Pedersen, G. Berardinelli, F. Frederiksen, P. Mogensen and A. Szufarska, "A flexible 5G frame structure design for frequency-division duplex cases", in IEEE Communications Magazine, vol. 54, no. 3, pp. 53-59, March 2016.

[19] R. Ferrús, O. Sallent, G. Baldini and L. Goratti, "LTE: the technology driver for future public safety communications", in IEEE Communications Magazine, vol. 51, no. 10, pp. 154-161, October 2013.

[20] 3GPP TS 36.331 V14.2.0, "Evolved Universal Terrestrial Radio Access (E-UTRA); Radio Resource Control (RRC); Protocol specification (Release 14)", March 2017.

[21] 5G-EmPOWER, available at: *http://empower.create-net.org/.*

[22] 5G-EmPOWER REST API Documentation, available at: *https://github.com/5g-empower/5g-empower.github.io/wiki/REST-API-documentation.*

[23] R. Cohen, G. Grebla, "Joint Scheduling and fast Cell Selection in OFDMA Wireless Networks", IEEE/ACM Transactions on Networking, vol. 23, no. 1, pp. 114-125, 2015.
[24] H. Kelleser, U. Pferschy, D. Pisinger, "Knapsack Problems", Springer-Verlag Berlin Heidelberg, 2004.
[25] R. Cohen, L. Katzir, D. Raz, "An Efficient Approximation for the Generalized Assignment Problem", Information Processing Letters, Elsevier, vol. 100, no. 4, pp. 162-166, Nov. 2006.
[26] T.D. Novlan, R. Krishna Ganti, A. Ghosh and J.G. Andrews, "Analytical Evaluation of Fractional frequency reuse for OFDMA Cellular Networks", IEEE Transactions on Wireless Communications, vol. 10. No. 12, pp. 4294-4305, Dec. 2011.
[27] M. Haenggi and R.K. Ganti, "Interference in Large Wireless Networks", Foundations and Trends in Networking, vol. 3, no. 2, pp. 127-248, 2009.
[28] M. Belesioti, I. Chochilouros, (editors), "System Use Cases and Requirements", Deliverable D2.1 of 5G-PPP SESAME project, December, 2015.
[29] J. Pérez-Romero, O. Sallent, R. Ferrus, R. Agustí, "Admission Control for Multi-tenant Radio Access Networks", ICC Workshop on Smart Communication Protocols and Algorithms, May, 2017.
[30] I. Giannoulakis (editor), "SESAME Final Architecture and PoC Assessment KPIs", Deliverable D2.5 of SESAME, December, 2016.
[31] 3GPP TS 36.314 v13.1.0, "Evolved Universal Terrestrial Radio Access (E-UTRA); Layer 2 – Measurements", March, 2016.
[32] 3GPP TS 32.425 v14.1.0 "Performance Management (PM); Performance measurements Evolved Universal Terrestrial Radio Access Network (E-UTRAN)", December, 2016
[33] Broadband forum "TR-069 CPE WAN Management Protocol", November, 2013.
[34] ETSI GS MEC 003 v1.1.1 "Mobile Edge Computing (MEC); Framework and reference architecture", March, 2016
[35] 3GPP TR 36.814 v9.0.0, "E-UTRA; Further advancements for E-UTRA physical layer aspects (Release 9)", March 2010.
[36] 3GPP TR 36.942 v12.0.0, "Radio Frequency (RF) system scenarios", September, 2014.
[37] P. Muñoz, O. Sallent, J. Pérez-Romero, "Capacity Self-Planning in Small Cell Multi-Tenant 5G Networks", IFIP/IEEE International Symposium on Integrated Network Management - 2nd International Workshop on Management of 5G Networks, Lisbon, May, 2017.
[38] 3GPP TS 32.450 v13.0.0, "Key Performance Indicators (KPI) for Evolved Universal Terrestrial Radio Access Network (E-UTRAN): Definitions", January, 2016.
[39] A. Whitehead (editor), "CESC Prototype design specifications and initial studies on Self-X and virtualization aspects", Deliverable D3.1 of 5G-PPP SESAME project, June, 2016.
[40] D. López-Pérez, M. Ding, H. Claussen and A.H. Jafari, "Towards 1 Gbps/UE in Cellular Systems: Understanding Ultra-Dense Small Cell Deployments", in IEEE Communications Surveys & Tutorials, vol. 17, no. 4, pp. 2078-2101, Fourthquarter 2015.