RAN Orchestration: A New Approach to Spectrum Management in Multi-Tenant 5G Networks

Shah Nawaz Khan[†], Leonardo Goratti[‡], Shahriar Hasan[§], Roberto Riggio[†]
[†] FBK CREATE-NET, Via Alla Cascata 56/D, 38123 Trento, Italy. {s.khan, rriggio}@fbk.eu
[‡]TriaGnoSys GmbH. Zodiac Inflight Innovations, D-82234 München, Germany. leonardo.goratti@zii.aero
[§]University of Trento, 38123 Trento, Italy. shahriar.hasan@studenti.unitn.it

***Abstract*—5G networks will incorporate new innovative technologies and concepts such as network virtualization, SDN/NFV, multi-tenancy and network slicing. Moreover, resource orchestration play a pivotal role to dynamically deploy network services and allocate resources. Orchestration is a control function for resource management in the network core and centralized cloud infrastructure. However, for end-to-end slicing and resource management, the orchestration functions must also be realized at the network edge i.e., the RAN segment. In this paper, we present an architecture for active RAN resource orchestration in which radio resource allocation to different tenants is dynamically scaled in real-time. A parallel can be drawn to the classic spectrum sharing scenarios as we evaluate the well known co-primary sharing model in a multi-tenant RAN context. Moreover, we extend SimuLTE, a well-known system level simulation model to integrate our RAN orchestration architecture and implement different scheduling policies. We evaluate system level network performance using fine-grained radio resource sharing approach and evaluate the involved performance-fairness trade-offs.**

I. Introduction

Performance targets set for 5G networks are very ambitious and will be difficult to reach with mere optimizations. Therefore, 5G networks are anticipated to incorporate new innovative technologies such as network virtualization, SDN, NFV, multi-tenancy and resource slicing. For radio coverage, additional radio frequency (RF) resources will be needed in different bands to address coverage and capacity concerns. However, the abundant literature on Cognitive Radio (CR) [1] indicates spectrum non-availability and highlights the need for radio resource sharing. The CR networks also face fundamental challenges that limit their practical realization and any implementations have been coarse-grained or non-overlapping [2]. In 5G networks however, resource sharing, multi-tenancy and network slicing are core functional considerations allowing for more active resource sharing options including radio resources. In this paper, we consider a 5G networks scenario that bases virtualization, multi-tenancy and resource sharing as core building blocks, all under the control of a Management and Network Orchestration (MANO) umbrella. While the traditional MANO functions are generally considered in the virtualized 5G core segment, we extend the orchestration functions to the RAN and present an active approach to radio resource orchestration. We consider a multi-tenant RAN scenario where tenants take a specific slice of radio resources and manage them separately for their end-users. The radio resource orchestrator facilitates the exchange of these resources by actively scaling in or out the amount of radio resources allocated to the deployed RAN tenants thereby realizing spectrum sharing in an active manner. We investigate a well-known co-primary spectrum sharing model in which a common pool of radio resources is available for use to a number of network operators. We evaluate our radio resource orchestration architecture using a modified SimuLTE model [3] that supports multi-tenancy in the RAN with slice isolation. The rest of the paper is organized as follows. Section II presents the related work on co-primary spectrum sharing. Section III presents the system architecture, control elements and scheduling policies developed for co-primary sharing. A modified SimuLTE model and evaluation resuts are detailed in section IV with a conclusion and summary in section V.

II. Related Work

Multi-tenancy and network slicing are core 5G networks concepts that are managed by MANO frameworks and enabled by technologies such as NFV and SDN [4] [5]. However, the adoption of SDN, NFV and MANO functions has been concentrated at the network core where centralized cloud infrastructures are used to support multi-tenant virtual network services. Multi-tenancy and dynamic resource provisioning using MANO control frameworks bring benefits to the infrastructure owners who can deploy, scale and manage virtual networks/slices and get the maximum benefit from the infrastructural resources. However, extension of multi-tenancy, resource sharing and MANO functions to the network edge i.e., the RAN segment, has not received the attention it warrants. Instead, only the traditional models of resource sharing have been considered where multi-tenancy manifests in the form of non-overlapping coverage or RAN infrastructure [6]. CR networks have

provided the foundation for most of the research done on spectrum sharing where different roles and privileges are attributed to the involved networks. Most common of these roles are primary and secondary user networks where secondary users are allowed opportunistic access to the primary users radio resources. This brings a high degree of uncertainty to the achievable throughput in secondary networks and poses several challenges such as spectrum sensing and interference avoidance; limitations that are well investigated [2]. However, several useful models of spectrum sharing have emerged such as mutual-renting and co-primary sharing which can be realized in multi-tenant 5G RAN context under an orchestrator's control. The co-primary spectrum sharing model brings a common pool of radio resources for shared access among several network operators [7]. Many research works have evaluated the co-primary sharing model and its benefits under the classic spectrum sharing context providing sufficient push for its consideration in 5G networks under a cooperative and coordinated manner [8] [9]. Building on our previous work that focused on fine-grained spectrum sharing using mutual renting model [11], in this work, we focus on extending the MANO domain to the network edge, supporting radio resource sharing using hierarchical control functions. We consider radio resource orchestration as a new realization of spectrum sharing models in 5G networks with specific focus on co-primary sharing model. To the best of our knowledge, most works on co-primary sharing in 5G context address interference mitigation, access coordination and policies for shared access. In the context of 5G however, RAN sharing scenarios are anticipated to manifest in the form of shared cells that are used by multiple operators with individual slices of radio resources. We focus here on very fine-grained spectrum sharing, in order of milliseconds and individual resource blocks.

III. Radio Resource Orchestration

The main assumption of our work is a 5G network that supports multi-tenancy in an end-to-end fashion. We assume an infrastructure provider having resources in the core and RAN that deploys virtual networks such as Mobile Virtual Network Operators (MNVOs) on the shared infrastructure and allocates them dedicated resource slices. These assumptions are well grounded considering that these features will be core part of 5G networks release 15 and beyond.

A. System Architecture

Figure 1 shows our considered radio resource orchestration architecture. At the top of hierarchy, control applications interact with a centralized control plane to configure network management parameters.For RAN, the control plane exposes parameters to the radio resource orchestrator (RRO) which uses them to configure

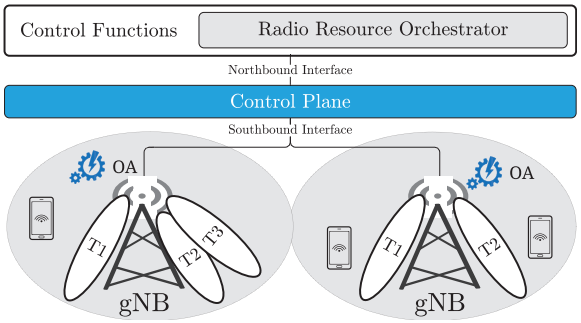

Fig. 1. System architecture and control elements

RAN elements. The RRO takes network state information and its domain knowledge such as pre-defined sharing policies to define the high level, non time-critical rules of spectrum allocation and sharing. This includes the allocation of bandwidth to new RAN tenants, cells where resources must be allocated and specifying the sharing rules. To enforce the RRO configurations at a more granular and time-critical level such as resource allocation and sharing decisions in individual cells or a set of few cells, orchestration agents (OA) are used. The OAs translate higher level directives into low-level control decisions which in this case, relate to bandwidth allocation to tenants and supporting co-primary resource sharing. At the OA level, the real-time network state is exposed through primitive parameters observed by the RAN elements. Together, the RRO and OAs control both non real-time and real-time radio resource management.

B. Control Flow

Figure 2 depicts the enumerated control flow among tenant schedulers, OAs and the RRO. We assume that the RRO keeps domain specific knowledge such as tenants' requirements, their eligibility for and amount of shared spectrum (1). This information may be given or derived from the tenants' service description files. When a new tenant is deployed, the RRO receives a trigger to specify the RAN segments in which it must be allocated, its dedicated bandwidth and whether the tenant is eligible for co-primary shared spectrum (2). Once deployed, a tenant acts like an isolated MVNO with full control over its allocated resources. We consider a communication interface between the tenants' schedulers and the OA which is used for two purposes. First, the OA has control over the tenants' bandwidth which it can dynamically change (i) if the RRO triggers it or (ii) if the co-primary shared spectrum is to be assigned to that tenant (3). Second, for fine-grained co-primary sharing detailed later, a tenants must expose its requirements for additional radio resources (4) which are allocated from the co-primary shared bandwidth. Depending on spectrum sharing policy, fairness among RAN tenants for co-primary shared bandwidth may not

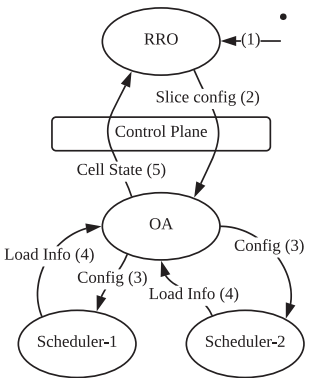

Fig. 2. Control flow among tenants, RTC and C3.

be guaranteed. Therefore, the OA regularly sends cell resource utilization statistics to the RRO (5). The RRO keeps a log of the resource utilization per tenant and uses it for balancing resource utilization among RAN tenants and modifying the time and frequency domain granularity. The time domain granularity determines the amount of time a particular tenant accesses the shared bandwidth. The frequency domain granularity refers to the bandwidth of acquired shared resources. In LTE networks, considered here only to make the description specific, the Physical Resource Blocks (PRB) is the most granular level of resource sharing in both time and frequency domains (0.5ms and 180kHz). In 5G-NR, there may be the possibility to have even more granular levels of resource sharing. Scheduling in LTE networks is done every 1ms (called Transmit Time Intervals; TTI) and the total PRBs are defined by the frequency bandwidth. In this context, the RRO defines time domain granularity using a window size w parameter. The w value serves as upper bound i.e., the maximum time a tenant can use the shared bandwidth. In frequency domain, RRO does not specify any limit as it is intrinsic in the available shared bandwidth N_{PRB}. However, the RRO does specify polices that affects the frequency domain granularity over time.

C. OA Scheduling Policies

Scheduling concerns finding a compromise between fairness and extracting the maximum value from resources. The OA modules perform real-time scheduling of the shared spectrum. When the RRO sends configuration for w and N_{PRB}, the OA take an active approach to allocating shared spectrum. The RRO is independent in its decisions for w with only the lower bound constraint of 1 TTI in LTE networks. The OAs use two distinct scheduling policies called Demand Aware Scheduling (DAS) and Preemptive Load-aware Scheduling (PLS). An OA communicates via its interface with tenant scheduler to get the tenant's specific load information. In DAS scheme, w is kept fixed while the frequency domain granularity is a variable in $1 \leq N_A \leq N_{PRB}$ range, decided by the OAs in real-time. At start, tenants are sorted according to their priorities if specified by RRO. Priority is an identifier that affects the shuffling of tenants in the scheduling queue after being scheduled. Assuming equal priority, in first window w_0, tenant at the head of queue is scheduled with entire co-primary shared bandwidth and is shuffled to the back. Subsequently, in each window w_i, a tenant T_x computes its mean downlink PRB requirement μ from the transmit buffers as a metric for real-time load and exposes it to OA. Considering that w_i is in order of milliseconds, the mean load approximation becomes an accurate estimate for the next interval w_{i+1}. The OA receives this information from tenants and formulates the scheduling list for window w_{i+1}. If the exposed request μ of the tenant at the head of queue is less than the dedicated resources of that tenant, it is shuffled one place back and is thereby excluded for scheduling in w_{i+1}. If however, the tenant at the head of queue is partially overloaded, a subset $1 \leq N_A < N_{PRB}$ of the shared resources is allocated and the tenant is shuffled to the back. Since real-time load is a random value, the DAS scheduling does not ensure fairness intrinsically among the RAN tenants. In essence, a tenant that is not loaded beyond its dedicated bandwidth will not be scheduled for the shared resources whereas a loaded tenant is scheduled more frequently. In order to compensate for this, the OA keeps a numeric credit value for each time the tenant at the head of the queue is shuffled without allocation. Partial allocation to a tenant also needs to be compensated for fairness and therefore, the OA increments its credit value proportional to the percentage of shared resources the tenant did not utilize. For example, if a tenant only gets 50 percent of the shared resources in its scheduling window, its credit is incremented by half a point compared with a tenant that did not use any additional resource in its scheduling window. It may happen that one tenant is exceedingly loaded while another is not loaded over a long period of time. In such scenario, the OA may not be able to compensate the unloaded tenant. For this case, the OA sends regular updates to RRO which can address this issue at higher abstraction level by e.g., changing the priority of tenant in other segment of the network. Allocation to potentially more than a single tenant per sharing window w_x aims to reduce radio resource wastage primarily in the frequency domain. Fundamentally, the DAS scheduling aims to get the maximum benefit from the co-primary shared resources by only allocating it to tenants that need it. The PLS is an inverted scheduling scheme compared with the DAS scheme. In PLS, the frequency domain granularity is kept fixed equal to the total bandwidth available for co-primary spectrum sharing N_{PRB} while the time domain granularity is a variable in $1 \leq N_{tti} \leq w$ range. The PLS

scheme is also different in terms of the information that is required at OAs. Each tenant only sends a binary signal representing a request for additional resources rather than exposing its real-time average PRB requirement. Moreover, since the time domain granularity is defined by TTI, the OAs can interrupt the shared resource assignment within w if the binary signal turns to zero. This enables the PLS scheme to avoid wastage of shared spectrum in the time domain. The PLS scheme also does not target fairness and uses the same compensation measures. For evaluations, we also implemented two well-known scheduling policies aimed at fairness and efficient resource utilization respectively. The first and most widely known scheduling approach is Round Robin (RR) in which the controller iterates over the RAN tenants and allocates the entire shared bandwidth. The duration of allocation is always equal to w and a constant waiting time is guaranteed. in RR, the tenants get a predictable level of benefit from the shared resource over time and no communication is needed between a tenant scheduler and OA. We also implemented a Load-Aware Round Robin (LA-RR) scheme in which the tenants expose their need for additional spectrum in order to be considered for allocation. If a request is not exposed, the tenant is always shuffled to the back of the queue. LA-RR filters down to RR if all tenants make requests all the time. The LA-RR does not guarantee fairness among RAN tenants as in DAS and PLS. All the scheduling polices that require information exchange between the tenant and OAs assume that the tenants are fair in their requests. A tenant can be selfish and expose incorrect requests for example to reserve its share of shared spectrum in case its requirements are predictable. The OA does not validate the tenants requests and we have not considered such cases.

IV. Simulation Analysis & Results

A. Modified SimuLTE Model

For this subsection, a basic understanding of LTE protocol stack and OMNeT++ Simulation platform is assumed. We use an LTE model since the 5G-NR standardization is ongoing and there a dearth of open source simulation platforms for analyzing new 5G networks concepts such as multi-tenancy, network slicing and resource sharing. For evaluation of our radio resource orchestration and RAN multi-tenancy architecture, we have modified an existing system level simulation model to add the required features. SimuLTE is a well known, system level simulation platform for LTE/LTE-Advanced networks analysis in OMNeT++ Simulator [3] [10]. This model has been used for system level analysis in several research articles but lacks the RAN multi-tenancy support as required for the work presented in this paper [11]. Moreover, to integrate the OA modules inside the simulated eNodeB (eNB) and RRO at

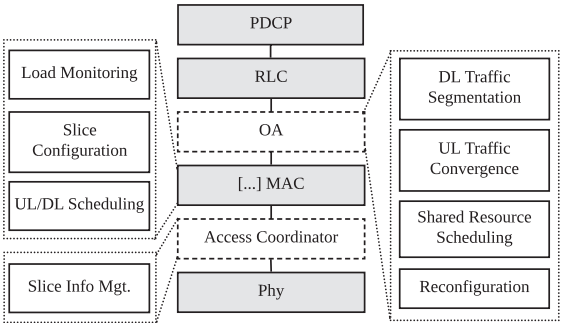

Fig. 3. Modified SimuLTE protocol stack for multi-tenant RAN

the network core, new OMNeT++ modules have been developed. Each RAN tenant in this model utilizes its dedicated radio resources and can share them with other tenants if spectrum sharing is enabled. Most of the modifications and updates have been done to the eNB compound OMNeT++ module of the SimuLTE model. This compound module represents a gNB with multiple RAN tenants as depicted in figure 1. Figure 3 shows the internal simple and compound modules of the modified eNB supporting the multi-tenant RAN feature. Instead of providing a complete LTE protocol stack for each tenant, we have implemented the orchestration agent module inside the LTE protocol stack. This means that a particular tenant is transparent to the network above Medium Access Control (MAC) and OA layers. The user equipments (UEs) data flows of all RAN tenants share the back-haul links and the core network. However, each tenant gets a dedicated MAC module that combines the main functions of slice-specific UEs uplink and downlink traffic scheduling, collecting the slice-specific load information, and reconfiguration of the radio resources based on the OA signals. The access coordinator is a utility module that maps the traffic of tenants to the wireless signals emitted by an eNB in the downlink direction and segments the UEs traffic to the correct MAC module in the uplink direction. The OA module implements the scheduling policies detailed above for co-primary spectrum sharing and interacts with tenant-specific MAC modules. The RRO (not shown in figure 3) allocates dedicated radio resource slices to the tenants which effectively creates a tenant-specific MAC module in the modified eNodeB stack. Table I shows the main simulation parameters considered in our evaluation. We simulated an Urban macro cell in 2100MHz band with 4 LTE based RAN tenants, each having 5MHz (25 PRBs) dedicated bandwidth and the OAs configured with additional 10MHz bandwidth available as common pool for co-primary spectrum sharing among these four tenants. The UEs of each tenant move with Random Waypoint mobility model and run different sessions of video streaming application that download varying sized videos from a server. The number of UEs per

TABLE I
CONFIGURATION PARAMETERS FOR SIMULATION ANALYSIS

	Parameter	Value	Parameter	Value
General	No. of Tenants	4	UE Mobility	RandomWP
	Sim Area	1 Sq Km	Geography	Urban Macro
PHY Layer	eNB Tx Power	10W	Frequency	2100MHz
	Shared Bandwidth	10 MHz	Slice Bandwidth	5MHz
	Antenna config	Omni-directional	eNB Height	25M
MAC Layer	Slice Scheduler	MaxCQI, PF	Queue Size	1MB
Application	Apps per UE	1	UE Application	Video Stream
	App Packet Size	1500B	Pkt Send Interval	1-3ms

tenant, their mobility and video streaming applications create different load fluctuations in the tenants' dedicated radio resources. The dedicated resources of the tenants are scheduled independently from the scheduling policy used by the OAs. In our simulations, the tenants scheduler uses Maximum Channel Quality Indicator (CQI) or Proportional Fair scheduling algorithms provided by the SimuLTE model. The orchestration agent module shown in figure 3 implements the four scheduling policies detailed previously to schedule the 10MHz co-primary shared spectrum among the simulated RAN tenants.

B. Simulation Results

Figure 4 shows the tenant specific load variation, expressed as downlink PRB utilization, for the four simulated tenants in their dedicated slice bandwidth of 5MHz. As radio resource scheduling in happens every 1ms in all tenant slices, the resulting curves for PRB allocation in each slice are difficult to interpret for load variations. To smooth out these fluctuations and improve readability of the figure of resource block allocation, we have applied a sliding window average function to all the four curves. As evident in figure 4, tenant-1 (T1) is the least loaded slice with average resource consumption mostly below its dedicated bandwidth. However, it should be noted that even though the T1 specific red curve never touches the maximum dedicated bandwidth (25 PRBs), it does not imply that T1 never needs any additional resources during the course of the simulation runs. In fact there are numerous instances where the load in T1 touches the maximum 25 PRB but those instances have been removed by the average function. The other three tenants are more loaded and their average requirement for radio resources exceeds their dedicated bandwidths thereby indicating frequent need for additional resources. Figure 4 is a reference performance indicator for the other results shown in the section as no co-primary shared resources are allocated in this instance. Figure 5 shows the same information of downlink resource block allocation when the co-primary spectrum is available to the tenants and the orchestration agent uses the four scheduling policies detailed previously. The figure indicates that all tenants gain benefit from the granular scheduling of the shared bandwidth in all

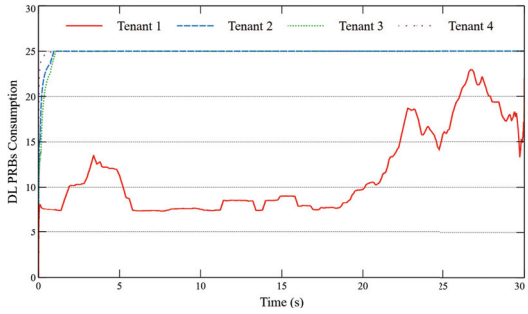

Fig. 4. Average downlink PRB consumption without shared resources

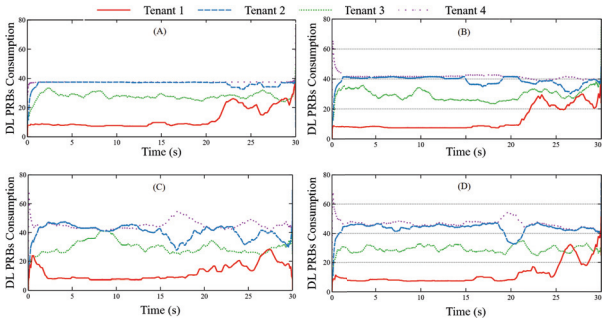

Fig. 5. Average downlink PRB consumption with shared resources

four scheduling policies which is an expected outcome. However, we also see the effect of the OA scheduling policy on the effective benefit for the involved RAN tenants. In figure 5 (A), the RR approach is used which ensures parity between the two most loaded tenants T3 and T4. All the loaded tenants see an increase in their downlink resource allocation due to the availability of additional spectrum and the fairness ensured by the RR policy. The LA-RR in figure 5 (B) trades off some fairness for better value from the co-primary shared spectrum and we see an increase in all curves compared with RR. However, the figure shows that both DAS and PLS policies (figure 5 C and D respectively) achieve the highest benefit from the shared resources with PLS being the best. The preemptive behavior of PLS scheme interrupts resource wastage in most fine-grained manner and achieves the maximum benefit for all four RAN tenants. Based on this result, it can be argued that at the cell-specific controller level, the maximum benefit achievable from the shared resource should be targeted and that the unfairness can be compensated at a higher abstraction such as at network level ensured by the RRO module. The impact on average cell throughput (joint application level throughput of all four tenant UEs) is shown in figure 6 together with the impact on downlink resource block allocation of all tenants. The average cell throughput indicates the joint benefit achieved from the shared resources under the same conditions using the four scheduling policies used by OAs. In conformance with the previous results, the DAS and PLS achieve the

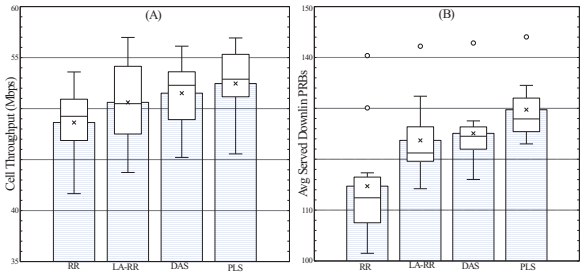

Fig. 6. Resource allocation policies

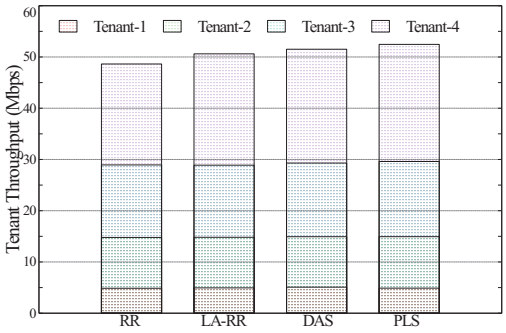

Fig. 7. Cell and tenant specific application layer throughput

best performance in both these indicators with the PLS scheduling policy achieving the best performance. These results have been collected across twenty runs for each policy case. As evident from this result, the joint cell throughput is maximized by the presented DAS and PLS policies with the main compromise made on fairness at the cell level. As an indication of the unfairness, figure 7 shows a segmented view of the average cell throughput as the sum of the tenant-specific slice throughput. It can be seen that even though DAS and PLS maximize the average cell throughput, the benefit is mainly obtained by T3 and T4 which are the most loaded slices as indicated in the reference figure. The least loaded slices of T1 and T2 get negligibly affected by the policies used for co-primary spectrum sharing as their benefit is constrained by their infrequent requests for additional resources. This indicates that at the granularity level considered in this work, fairness among RAN tenants as targeted by RR and LA-RR does not translate into a quantifiable benefit. On the other hand, DAS and PLS inherently achieve fairness as the load variation among RAN tenants decrease i.e, when all tenants are loaded, the DAS and PLS converge on RR policy. Any considerable difference between tenants such as between T1 and T4 can be compensated for, using network level control modules such as the RRO in our presented radio resource orchestration architecture.

V. Conclusion & Future Work

We presented RAN orchestration as a new approach to realizing spectrum sharing in multi-tenant 5G networks. We presented a hierarchical RAN orchestration architecture that can be applied to schedule shared spectrum in real-time, fine-grained manner. We presented a the orchestration architecture with control elements at different abstraction levels in the network. Two scheduling polices were presented and analyzed in LTE based RAN with a granularity level of milliseconds in time domain and PRBs in frequency domain. Results were presented with a discussion on the achievable throughput benefits and the fairness consideration among the RAN tenants. The presented fine-grained radio resource sharing approach requires a tighter integration between the orchestration agents and tenant-specific resource scheduler in order to expose the required load information. In future work, we plan to investigate approaches to dynamically scale the time and frequency domain granularity in different segments of the network to address the fairness issues at the network level. We also plan to integrate heterogeneous access technologies evaluate fine-grained resource sharing in heterogeneous networks.

Acknowledgment

Research leading to these results received funding from the European Unions H2020 Research and Innovation Action under Grant Agreement H2020-ICT-644843 (5GESSENCE Project)

References

[1] P. Setoodeh, and S. Haykin, “Cognitive Radio Networks,” in Fundamentals of Cognitive Radio , Wiley Telecom, 2017
[2] M. Mustonen, M. Matinmikko, D. Roberson and S. Yrjölä, “Evaluation of recent spectrum sharing models from the regulatory point of view,” 1st International Conference on 5G for Ubiquitous Connectivity, Akaslompolo, 2014, pp. 11-16
[3] SimuLTE: ”An Open Source LTE User-Plane Simulation Model”, Online: http://simulte.com/. Accessed: 2018
[4] ETSI, ”Open Source MANO”, Online: http://www.etsi.org/
[5] A. Gupta and R. K. Jha, ”A Survey of 5G Network: Architecture and Emerging Technologies,” in IEEE Access, vol. 3, pp. 1206-1232, 2015
[6] L. Zhang, M. Xiao, G. Wu, M. Alam, Y. C. Liang and S. Li, ”A Survey of Advanced Techniques for Spectrum Sharing in 5G Networks,” in IEEE Wireless Communications, vol. 24, no. 5, pp. 44-51, October 2017
[7] P. Luoto, M. Bennis, P. Pirinen, S. Samarakoon and M. Latva-aho,“Enhanced Co-Primary Spectrum Sharing Method for Multi-Operator Networks,” in IEEE Transactions on Mobile Computing, vol. 9, no. 99, pp. 1-1 2017
[8] B. Singh, K. Koufos, O. Tirkkonen and R. Berry, ”Co-primary inter-operator spectrum sharing over a limited spectrum pool using repeated games,” in IEEE International Conference on Communications (ICC), London, 2015, pp. 1494-149
[9] B. Cho, K. Koufos, R. Jntti and S. L. Kim, ”Co-Primary Spectrum Sharing for Inter-Operator Device-to-Device Communication,” in IEEE Journal on Selected Areas in Communications, vol. 35, no. 1, pp. 91-105, Jan. 2017
[10] Andras V.: OMNeT++: Discrete Event Simulator. Online: http://omnetpp.org: Accessed: July 9th, (2017)
[11] S. N. Khan, L. Goratti, R. Riggio, and S. Hasan, “On Active, Fine-Grained RAN and Spectrum Sharing in Multi-Tenant 5G Networks”, in Proc IEEE International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC), Montreal Canada 2017