Received February 18, 2015, accepted March 9, 2015, date of publication April 13, 2015, date of current version May 5, 2015.

*Digital Object Identifier 10.1109/ACCESS.2015.2422266*

# Cooperative Radio Resource Management in Heterogeneous Cloud Radio Access Networks

**MIKHAIL GERASIMENKO[1], DMITRI MOLTCHANOV[1], ROMAN FLOREA[1], SERGEY ANDREEV[1], YEVGENI KOUCHERYAVY[1], NAGEEN HIMAYAT[2], SHU-PING YEH[2], AND SHILPA TALWAR[2]**

[1]Department of Electronics and Communications Engineering, Tampere University of Technology, Tampere 33720, Finland
[2]Intel Corporation, Santa Clara, CA 95054, USA

Corresponding author: S. Andreev (sergey.andreev@tut.fi)

This work was supported in part by Intel Corporation and in part by the Internet of Things Program of DIGILE through Tekes. The work of S. Andreev was supported by the Academy of Finland with a Post-Doctoral Researcher grant and by Nokia Foundation with a Jorma Ollila Grant.

**ABSTRACT** Responding to the unprecedented challenges imposed by the 5G communications ecosystem, emerging heterogeneous network architectures allow for improved integration between multiple radio access technologies. When combined with advanced cloud infrastructures, they bring to life a novel paradigm of heterogeneous cloud radio access network (H-CRAN). The novel H-CRAN architecture opens door to improved network-wide management, including coordinated cross-cell radio resource allocation. In this paper, emphasizing the lack of theoretical performance analysis, we specifically address the problem of cooperative radio resource management in H-CRAN by providing a comprehensive mathematical methodology for its real-time performance optimization. Our approach enables flexible balance between throughput and fairness metrics, as may be desired by the network operator, and demonstrates attractive benefits when compared against the state-of-the-art multiradio resource allocation strategies. The resulting algorithms are suitable for efficient online implementation, which principal feasibility is confirmed by our proof-of-concept prototype.

**INDEX TERMS** Heterogeneous network, cloud infrastructure, heterogeneous cloud radio access network, cooperative radio resource management, mathematical methodology, prototyping.

## I. INTRODUCTION AND MOTIVATION

We are rapidly moving to the 5G era, where everything that can benefit from a wireless connection will become a part of next-generation network infrastructure. In these exciting times, when traffic from wireless devices is expected to exceed data from wired equipment and the overall mobile traffic demand might increase 11-fold within only 5 years from now, we are also facing unprecedented challenges to make this vision come true. Accordingly, the networks of tomorrow will need to reach the staggering densities of devices and network infrastructure nodes, harness very high carrier frequencies with emerging millimeter wave (mmWave) technologies, and support extreme numbers of antennas in massive multiple-input multiple-output (MIMO) installations [1]. Additional spectral resources will have to be made available together with improved levels of intelligence and flexibility across prospective 5G deployments, increasingly mindful of power and cost efficiencies.

Responding to these challenges, the paradigm of heterogeneous network (HetNet) has recently emerged as advanced networking architecture comprising a hierarchy of 3GPP LTE macro cells for ubiquitous coverage and connectivity enhanced by small cells of different sizes and across various radio access technologies (RATs) to augment capacity. These small cells may reside in both licensed and unlicensed spectrum offering open, closed, or hybrid user access [2] and include pico and femto cells, WiFi and WiGig access points, remote radio heads and relay nodes, integrated WiFi-LTE small cells, etc. Recent progress in 3GPP standardization allows to efficiently coordinate between such alternative radio access networks (RANs) to unlock substantial gains in network capacity and user connectivity experience [3]. However, this improved coordination comes with a price, as macro and small cells have to be connected via low-latency high-rate backhaul links striving for the maximum flexibility of HetNet management for e.g., enhanced capacity, seamless mobility, and robust interference mitigation [4].

Depending on the effective backhaul restrictions, different levels of coordination within 5G HetNet architecture may become feasible. For example, if only *non-ideal* constrained backhaul is available to a mobile network operator, coordination via anchor-booster architecture may be employed where anchor macro base station (BS) provides

overall network management and diverse multi-radio small cells boost user data rates by enabling opportunistic traffic offloading (see TR 36.842). Alternatively, in case of *near-ideal* (e.g., optical fiber) backhaul with higher capacity and lower latency, the baseband signals from numerous low-power small cells may be received and processed at a remote centralized server platform. This attractive architecture, named Cloud RAN, becomes increasingly preferred by the network operators with prevailing fiber and inexpensive wireless fronthaul connections, primarily in ultra-dense HetNet deployments covering areas with high traffic demand. Today, when up to 80% of mobile operator's CAPEX is spent on the RAN, the concept of Cloud RAN allows to significantly lower capital/operational expenditures as well dramatically reduce energy consumption of wireless infrastructure.

In Cloud RAN, the remote radio head (RRH) unit, which is a simplified low-power node, utilizes the high-rate fronthaul links to compress and forward the baseband signals from mobile user equipment (UE) to the centralized base band unit (BBU) thus acting as soft relay. Therefore, fronthaul capacity constraints impose a fundamental limitation on the resultant system operation and advanced signal processing solutions [5] together with dynamic radio resource management [6] are required to maintain acceptable Cloud RAN performance. The available Cloud RAN capacity limits also depend (i) on the practical backhaul constraints [7] and respective optimization [8], (ii) on the utilized uplink RRH association strategies [9] with corresponding restrictions on implementation complexity and radio resource consumption [10], as well as (iii) on the employed decentralized beamforming algorithms [11] and large-scale distributed MIMO-aware power and antenna selection schemes [12].

Most recently, the concept of heterogeneous Cloud RAN (H-CRAN) has been proposed [13] as cost-efficient solution to further improve on the available cooperative gains in HetNets through their combination with the ''signal processing cloud''. Conveniently located at the intersection of heterogeneous networking and cloud computing, H-CRAN inherits the attractive benefits of both realms facilitating interference mitigation, scalability, and radio resource control with its cooperative processing and networking techniques. Correspondingly, radio resource management of low-power nodes (LPNs) is moved to a virtual BS, which is a part of the processing capacity allocated from the physical BBU pool on the cloud server. In summary, while technological features and core principles behind H-CRAN have been outlined towards opening path to commercial H-CRAN based 5G systems [13], major research challenges remain along the lines of theoretical performance analysis and optimal resource allocation to understand the ultimate potential of this promising innovation.

In this article, building on our knowledge of corresponding technology and latest developments in 3GPP standardization, we focus on the problem of cooperative radio resource management in 5G-grade H-CRAN systems by providing a comprehensive methodology for real-time performance optimization of H-CRANs. Our proposed solution allows to dynamically control the amount of resources allocated to the end users for two alternative metrics of interest, namely, the fairness of resulting resource shares across all the available RANs and the overall system throughput. We specifically concentrate on how to manage the dynamic H-CRAN systems by flexibly exploiting the trade-offs between these metrics. Further, we thoroughly compare performance of network-centric, network-assisted, and UE-centric resource allocation mechanisms in characteristic H-CRAN environment with different levels of available LTE/WiFi integration, as well as review our respective proof-of-concept developments.

## II. H-CRAN: TECHNOLOGY DEPLOYMENT AND STANDARDS BACKGROUND

The H-CRAN technology is still far away from the real deployments, but ongoing work within 3GPP standards is beginning to address coordinated use of multiple RATs as part of single operator managed multi-radio network. We continue with a concise review of the respective efforts.

In a loose coordination model, the 3GPP LTE Release 11, the access network discovery and selection function (ANDSF) concept is used to manage the interworking between WLAN and 3GPP networks (see TS 23.402) via the ANDSF policy server within the core network. Here, the operator is able to specify relatively static policies on discovery and use of WLAN resources within the network, leaving the task of network selection to the UE, which factors local operating environment and dynamically changing radio link conditions in its decisions. ANDSF-enabled architecture is shown in Fig. 1, subplot A.

However, the *UE-centric* decisions are sub-optimal, as the UE is unaware of the link conditions and radio resource requirements of other users sharing the radio network. Moreover, the multi-radio network nodes within the HetNet deployments possess very limited knowledge about each others' radio resource conditions and usage, which further reduces the efficiency of radio resource utilization in the network. On top of that, the mobility anchor between the WLAN and the 3GPP links resides within the core network (typically, at the P-GW), making traffic steering between WLAN and 3GPP links very expensive when adapting to dynamically changing link conditions.

Recently concluded (as part of Release 12) work on WLAN/3GPP radio interworking (see TR 37.834) attempts to partly alleviate the above issues. Through proprietary coordination within the RAN, the cellular BS (eNB or eNodeB) is able to set thresholds related to link quality and WLAN loading, which can facilitate traffic steering between WLAN and 3GPP RANs. However, the coordination required in the network to set the appropriate thresholds is not specified. Further, since the mobility anchor of WLAN/3GPP links still resides in the core network, it remains inefficient to make fast traffic steering decisions with this solution. In what

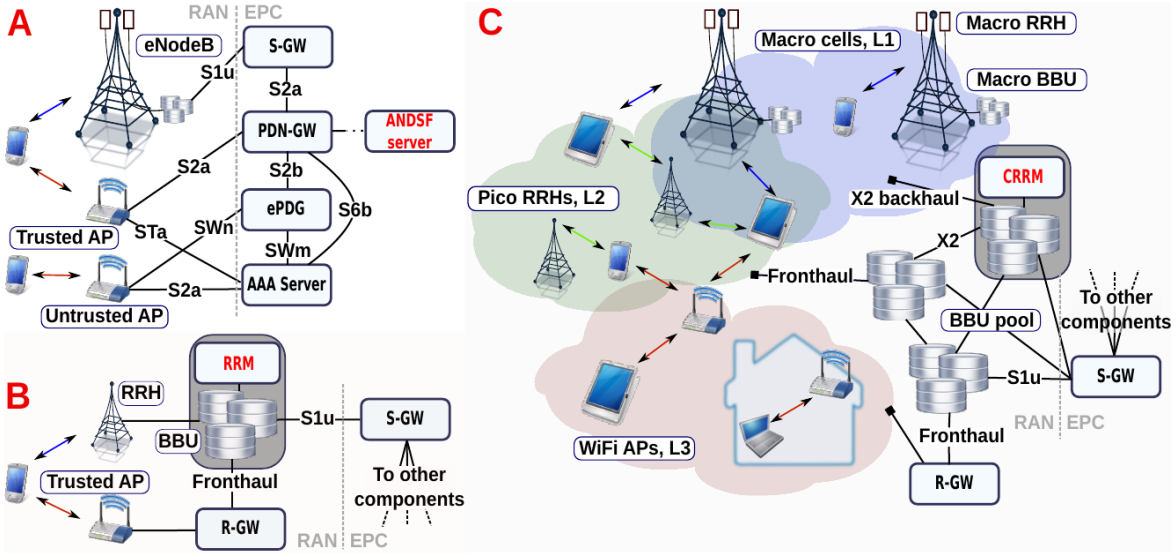



**FIGURE 1.** Example of H-CRAN deployment and system modes. (A) ANDSF-enabled architecture. (B) Anchor-booster architecture. (C) H-CRAN architecture.

follows, we term mechanisms based on this integration option as *network-assisted*.

Current efforts under way within 3GPP are targeting tighter integration of WLAN within the 3GPP RAN. Recent proposals (see e.g., RP-140685, RP-140738) aim to utilize WLAN as a secondary carrier, anchored at the eNB, within the 3GPP RAN. The proposed architecture extends the benefits of Release 12 dual connectivity anchor-booster system design introduced for 3GPP small cells (see TR 36.842), as well as the existing 3GPP carrier aggregation framework to also include non-3GPP RATs, such as the WLAN access example used for illustration here. It is expected that 3GPP may consider such architectures for Release 13 standardization.

It is also anticipated that as part of this work 3GPP will consider standardizing the interface between eNB and WLAN access points (APs) for non-collocated WLAN/3GPP deployments. Note that if adopted by 3GPP, this integrated *network-controlled* architecture will extend the benefits of LTE-based anchor-booster schemes and make the coordinated radio resource management also available for non-3GPP WLAN networks. Importantly, the use of the eNB as an anchor node for WLAN connections, allows users to employ the LTE network for control and management functions, leaving the WLAN capacity to be used solely for data offloading. The simplified architecture of this approach is shown in the Fig. 1, subplot B. In the plot, the radio resource management function is marked as RRM module, while the interface from WiFi to BBU is not standardized yet, the connection is assumed to be passed through a specific gateway (called RAN gateway or R-GW in the figure) which performs the interfaces matching.

While anchor-booster architectures allow for coordinated use of radio resources within the anchor cell (typically, macro cell) coverage area, the overall system performance can be improved further if radio resource coordination across anchor cells is also enabled. In practice, different approaches may be used to allow for such coordination. In the distributed model, eNB to eNB coordination may be achieved over the X2 interface. Alternately, a centralized radio resource controller may be used to manage system-wide radio resources. A 3GPP study is currently underway to explore such architecture for multi-RAT networks (see TR 37.870).

For deployments that can exploit high-rate fiber connections, a Cloud RAN architecture becomes feasible, which links RRH with simple functionality to centralized BBU pool within the cloud. 3GPP allows for such Cloud RAN deployments, but there is need for additional standardization efforts as such architecture collapses the entire RAN functionality within a single centralized node. Such emerging architectures can easily accommodate non-3GPP RRH nodes to allow for centralized multi-radio coordination.

The H-CRAN concept introduced in [13] and reviewed in Section I, is similar in principle to the multi-radio Cloud RAN discussed here, but also allows for centralized processing to occur within the EPC, when the nodes use the S1 interface to connect to the centralized server. As noted, the centralized network allows for coordinated/cooperative radio resource management, which will take into account not only interference issues, but also load variations (e.g., busy hour effect) and mobility of the UEs (e.g., high mobility UEs could be by default offloaded to macro cell).

In our further evaluation, we assume a H-CRAN deployment, which allows for centralized management of system radio resources with a dedicated entity, named Cooperative Radio Resource Manager (CRRM), whereas the connection to the CRRM server is done through the same X2 backhaul interfaces. Deployment and system model of this approach is shown in Fig. 1, subplot C.

## III. H-CRAN RESOURCE OPTIMIZATION WITH CRRM

### A. PROPOSED OPTIMIZATION METHODOLOGY

In this work, we consider H-CRAN environment with a number of multi-radio RANs, termed layers. Following the concepts outlined in Section II, all the radio access nodes on these layers are assumed to be connected to the CRRM, which is responsible for centralized cross-RAT resource allocation. A particular RAN typically features its individual set of lower-layer channel adaptation and signal processing mechanisms controlling radio links of associated users. These important parameters, along with those pertaining to user traffic demands and radio connectivity options, could be made available to the CRRM. Consequently, CRRM shall become responsible for the overall H-CRAN performance optimization. However, the development of appropriate optimization procedures and respective low-complexity real-time algorithms remains a very complex research problem. Alternatively, the role of CRRM may be limited solely to the traffic optimization functionality within H-CRAN. In this work, we thus specifically concentrate on such cooperative radio resource management methodology.

The system state of H-CRAN at any given instant of time can be described by the state of the traffic demands at individual nodes, their feasible RAN connectivity options, and geographic locations within the service area of interest. The information on the evolution of such metrics is reported by the UEs or, alternatively, by the radio access nodes to the CRRM. At the moments of state changes, CRRM assumes the values of these metrics as its input and optimizes the corresponding resource allocations across the available RANs. The optimal allocations are then advertised to the UEs via appropriate control interfaces and continue to be in effect until the next state change occurs. The frequency of such optimization depends on the dynamics of input parameters. As long as they remain unchanged, no update on the use of resources is needed.

### B. PERFORMANCE CRITERIA

There are several crucial performance criteria, which can be adopted when optimizing performance of H-CRAN environment. Among others, the overall system throughput, $T$, and the fairness of resulting resource allocations are of particular importance. For a given set of satisfied traffic demands, the latter can be characterized by the Jain's fairness index, $F$. The mobile network operator could also be interested in minimizing the energy consumption of mobile UEs and/or providing a certain degree of prioritization between the users. Subscriber priorities enable utilization of flexible pricing schemes, which open path to cost optimization. Various metrics can be combined when formulating the target optimization task.

Importantly, fairness of resource allocations is a user-centric metric, whereas system throughput is a network-centric criterion. In addition to the inherent trade-off between them, H-CRAN adds an extra degree of complexity. Wireless link-level mechanisms employed by the state-of-the-art communication systems force the UEs to use various channel adaptation parameters eventually resulting in different effective data rates over the same amount of radio resources. To maximize the overall system throughput, the H-CRAN resources must be assigned to UEs with maximum instantaneous spectral efficiency, which naturally contradicts the fairness requirement. Hence, a critical point for the operator is to have a flexible balance between these two metrics.

There exist two alternative fairness criteria resulting in different trade-offs between $T$ and $F$, known as max-min and proportional fairness. The max-min criterion delivers the maximum possible fairness for given conditions. Denoting by $M$ the number of demands and by $P_d$ the number of available data transmission paths for demand $d$ in H-CRAN, the respective objective is to maximize the allocations $h_d = \sum_{d=1}^{P_d} x_{dp}$ subject to the capacity constraints of RRHs. We say that $\vec{h}$ provides the max-min allocation if it is lexicographically maximal among all the feasible allocation vectors sorted in non-decreasing order. The resulting optimization task is classified as linear programming problem, thus making feasible efficient real-time implementations.

However, as discussed above, the max-min approach blindly trades the overall system throughput for fairness and is thus hard to manage. A possible workaround to this is with the alternative proportional fairness (PF) criterion penalizing the long-distant data paths more heavily and, hence, resulting in better performance of shorter paths. However, PF-centric optimization is of convex programming type, thus imposing additional constraints on the real-time implementation, especially in large-scale dense H-CRANs. Fortunately, the trade-off between system throughput and fairness in H-CRANs can be flexibly controlled with our proposed modified max-min criterion, still maintaining simplicity of on-line implementation.

### C. PERFORMING OPTIMIZATION IN H-CRANs

Consider a certain time instant $t$ when the change in the system state invokes the resource optimization function at CRRM. The corresponding resource allocation model incorporates three steps: (i) H-CRAN topology modeling, (ii) specification of the appropriate optimization model, and (iii) efficient solution algorithm. In H-CRAN uplink, we are interested in the so-called *bifurcated* resource allocation allowing a particular traffic demand to be split flexibly among the available radio interfaces. This corresponds to the case when a multi-radio UE may transmit on more than one radio simultaneously and results in relative simplicity of our resource allocation algorithm, as well as its core optimization routine, making them implementable in a practical CRRM.

The conventional approach to the aforementioned allocation problem is to maximize the minimum signal-to-interference-plus-noise ratio (SINR) across all the users. However, this may not be feasible due to a number of inherent system limitations. First, the resulting problem is of convex programming type, which is significantly more

computationally intensive than linear programming models. Second, in the presence of multiple RANs with different available bandwidths, the appropriate objective function takes a complex form. To reduce complexity, our proposed methodology is thus based on several clever approximations of the effective interference levels by the appropriate interference margins.

Another important assumption we need to adopt is that all traffic demands are greedy and elastic. Recall that greedy (full-buffer) traffic occupies all the allocated resource, while elasticity implies adaptiveness to the actually available resource allocations. With the above two assumptions, the considered H-CRAN system is fully characterized by the users' radio connectivity options at time $t$. As we will see next, the formulated model provides accurate performance results, while being extremely computationally efficient. As a consequence, it is suitable for the real-time performance optimization of large-scale and highly heterogeneous H-CRANs.

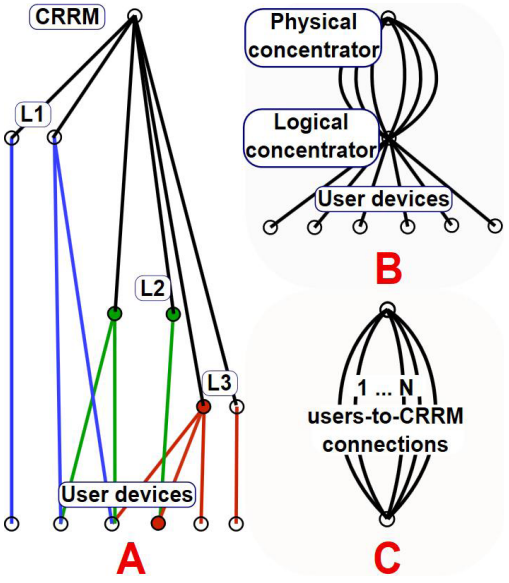



**FIGURE 2.** H-CRAN deployment analysis.

The principles of the proposed topology modeling are illustrated in Fig. 2. The original topology is outlined in subplot A, where colors highlight different types of radio access nodes. Assuming perfect dimensioning, the bottleneck of H-CRAN shifts to the radio interfaces. Therefore, here we can abstract away the links connecting radio access nodes with the BBUs, thus resulting in a graph shown in subplot B. Further, by removing an additional set of links connecting UEs with the radio access nodes, we represent generic H-CRAN topology as a two-node graph with one logical and one physical concentrators. The number of links connecting the concentrators equals to the number of radio access nodes at all layers. In this interpretation, we keep track of the users' connectivity by specifying the set of available paths between two nodes (subplot C).

Correspondingly, the capacity and demand constraints are formulated as:

$$\sum_{p=1}^{P_d} \alpha_{dp} x_{dp} = h_d,$$
$$\sum_{d=1}^{M} \sum_{p=1}^{P_d} \delta_{edp} x_{dp} = B_e, \quad (1)$$

where $\delta_{edp}$ is a $0-1$ variable specifying whether link $e$ is used to carry a part of demand $d$ over path $p$, $B_e$ is the bandwidth of radio access node $e$. The primary task is then to lexicographically maximize the vector of bandwidth allocations $\vec{h}_d$ leading to a linear programming problem.

In the specification above, it is natural to set the coefficient $\alpha_{dp}$ to the current spectral efficiency of mobile user $d$ over wireless interface $p$. However, the trade-off with this choice of coefficients will be in favor of fairness. To allow for a controllable balance between fairness and system throughput, we propose a special form of the spectral efficiency function. Defining $\alpha_{dp} = \beta^{s_{dp}}$, we can use parameter $\beta$ to adjust our optimization algorithm. Setting $\beta = 1$, we arrive at the classic max-min resource distribution in terms of fairness of allocated rates. With $\beta > 0$, users having higher spectral efficiency receive more rate leading to better system throughput. In the limit $\beta \to \infty$, the entire system bandwidth is assigned to the users with the best possible spectral efficiency, hence maximizing the system throughput.

### D. EXTENSIONS TO PROPOSED FRAMEWORK

The proposed optimization framework serves as a solid baseline that can be further extended to take into account additional metrics of interest. One potential extension is based on introducing additional coefficients to (1). For example, by adding coefficients $\gamma_d$, $d = 1, 2, \ldots, M$ to the first equation, we can take into account the priority of resource allocation across a certain number of classes. Further, by introducing extra weights corresponding to the energy consumption associated with different radio interfaces and/or spectral efficiencies, we might also optimize UE energy consumption across H-CRAN. In addition, the finite traffic demands can be incorporated into our model without any significant modification by specifying them explicitly. The core difference is that the resulting optimization may not have a solution when the demands exceed the total system capacity.

Our model can also be modified to accommodate different objectives and/or environments. The non-bifurcated allocation (when UEs cannot split their traffic) will make the optimization problem become of mixed-integer programming (MIP) type, thus significantly complicating the solution algorithm. Moreover, this problem is known to be NP-complete even for a single source and multiple destinations. However, in case of H-CRAN where all data paths are of length two, an efficient optimization algorithm may be feasible.

Modifying our objective function, we can also address the case of different routing costs over individual RANs. For example, higher routing cost over macro LTE network will force UEs to offload more traffic onto WiFi or small cell interfaces. Note that this modification does not render the resulting problem outside of linear programming framework. Targeting delay optimization in H-CRAN will also require the change of the objective function and would lead to the convex programming problem. Finally, with appropriate modifications, our model is suitable for downlink optimization as well.

## IV. NUMERICAL PERFORMANCE EVALUATION

### A. REPRESENTATIVE ARCHITECTURES AND DEPLOYMENTS

To rigorously evaluate our performance optimization solutions, we have conducted thorough analysis of several radio resource management schemes in realistic H-CRAN ecosystem by constructing a number of representative deployment scenarios. Along these lines, we employed our capable analytical environment, which has been initially calibrated with detailed system-level simulation (SLS) tool used extensively in our past research [3]. The discussed environment comprehensively characterizes the core performance of alternative resource allocation strategies by abstracting away less impactful features of practical deployments to reduce complexity.

Whereas some system-level features have been simplified in the present tool (e.g., inter-cell interference and fading variations are captured with appropriate margins), our employed analytical environment remains sufficiently accurate to provide the first-order understanding of the discussed concepts. With inter-cell and inter-layer interference being simplified, we concentrate our attention on the performance of a characteristic (typical) 3GPP LTE cell under the coverage of macro BS. Other infrastructure nodes, such as pico cells and WiFi access points, are deployed uniformly within the area of interest according to the current 3GPP specifications.

In what follows, we build three representative H-CRAN deployment scenarios corresponding to typical network operator strategies and respective LTE/WiFi integration choices: (i) operator deploys LTE macro network in licensed bands and also owns "carrier-grade" WiFi network in unlicensed spectrum; (ii) only LTE technology is employed by the operator, with a macro tier and an additional pico small-cell tier in separate licensed frequencies, and (iii) all of the above options are available to the operator, that is, LTE macro, pico, and WiFi RANs. Naturally, the latter option results in higher potential capacity and better expected flexibility, and we are interested to conclude on the extent of available benefits.

The densities of pico and WiFi small cells in our deployment scenarios may vary, but the target value of 4 radio access nodes per a macro cell is assumed to mimic contemporary urban conditions. The total number of UEs deployed in our system is fixed to 60 in accordance with the relevant 3GPP documents, and their distribution across the area of interest is also taken as uniform. As mentioned previously, backhaul and fronthaul links are assumed to have higher capacity than the radio links under consideration, and below we concentrate on evaluating the effects related solely to the radio channel capacity. All of the UEs are assumed to be static during such experiment, whereas their geographical positions change across different replications. Other important system parameters are summarized in Table 1. For brevity, both WiFi APs and LTE pico RRHs are named there low-power nodes (LPNs).

**TABLE 1. Primary deployment parameters.**

| Parameter | Value |
|---|---|
| LTE/WiFi configuration | 10 MHz FDD / 20 MHz |
| Layout | 1 macro cell, several LPNs |
| Macro/LPN-UE pathloss model | ITU UMa/UMi |
| Macro/LPN antenna gain | 17/6 dB |
| Macro/Pico/WiFi max. power | 43/30/20 dBm |
| LPN/UE max. power | 23/20 (LTE/WiFi) dBm |
| LTE/WiFi power control | Max power |
| UE/Macro/LPN antenna height | 1.5/25/10 m |
| UE noise figure/feeder loss | 5 dB/0 dB |
| Traffic model | Full-buffer |
| LPN/UE deployment type | Uniform |
| LPN/UE-Macro distance | $>$ 75/35 m |
| LPN/UE-UE distance | $>$ 40/10 m |
| Trials per experiment | 1000 |

### B. PERFORMANCE EVALUATION AND COMPARISON

In this subsection, we conduct performance comparison of prospective H-CRAN deployment (enabling full network control) with UE-centric and network-assisted multi-radio integration architectures having lower degrees of manageability. As an illustrative example of UE-centric solution we employ a simple "greedy" (max-usage) scheme, when UE attempts to utilize all the available radio resources on all possible RANs it may connect to. This intuitive strategy does not impose tight signaling requirements on neither network nor UE and performs reasonably well from the individual user throughput perspective. Most importantly, this option is very easy to implement in real equipment.

Another characteristic example, which is a network-assisted resource allocation scheme, is exploiting the concept of cell range expansion. It controls effective association thresholds of WiFi and LTE small cells for the users under macro cell coverage. This approach allows to include the effect of load variations in individual RANs by increasing/deceasing the number of UEs associated with the small cells. Naturally, this method also requires a separate decision-making entity implemented at eNB together with

the corresponding signaling to supply UEs with the relevant assistance information.

In our implementation of network-assisted solution, we consider a non-bifurcated option, which allows the UE to prefer a single RAN at a time. This modeling choice is valid in practice, as otherwise this scheme would lose to the UE-centric solution by allocating maximum resources to the UEs with good channel conditions and not even allowing the UEs with poor connectivity to attempt WiFi/pico cells. In other words, the short description of the implemented algorithm is as follows: the UE first attempts the WiFi layer, but if the resulting signal quality is bad it will then attempt the pico layer and, if there are no other feasible options, the UE will be served by the macro layer. Finally, we also consider the network-centric mechanism enabled by H-CRAN/CRRM, which centrally optimizes cross-RAT resource allocations based on the effective UE radio link quality. In fact, both network-assisted and network-centric mechanisms perform better in H-CRAN environment. However, as we demonstrate further, the centralized solution offers better control flexibility.

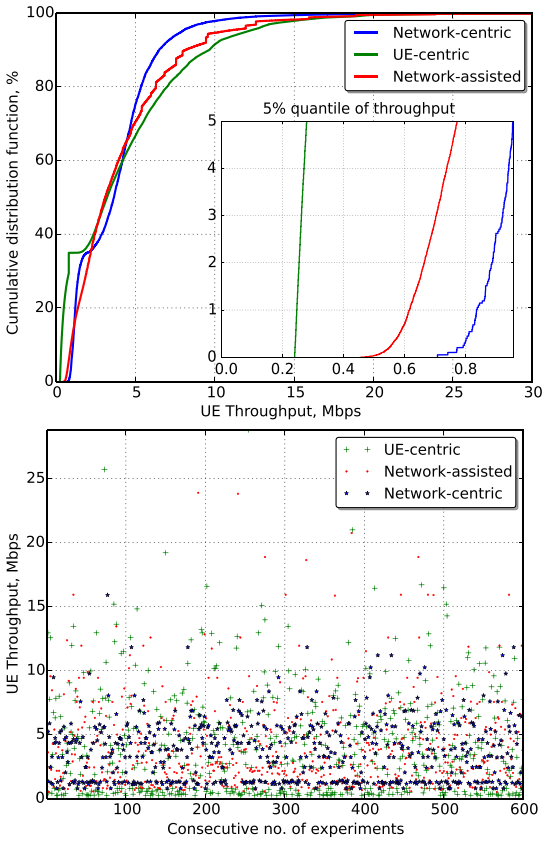



**FIGURE 3.** **Per-UE throughput results.**

In Fig. 3 (top subplot), all three schemes are compared in terms of per-UE throughput. The threshold for the network-assisted scheme for both LTE and WiFi small cells has been set to the maximum sensitivity level. In network-centric solution, CRRM takes as the input the vector of spectral efficiencies for each user first without the modifications discussed in Section III. As the result of this default configuration, all three solutions demonstrate comparable results in terms of the average throughput, but the network-centric mechanism performs much better at the cell edge.

The results for per-UE throughput are illustrated in the bottom subplot of Fig. 3 in the per-UE manner. From the figure we learn that the UE-centric scheme is characterized by lower fairness compared to the network-centric/assisted methods. This conclusion can be confirmed by studying the 5%-quantile sub-figure embedded in Fig. 3, top subplot. Another interesting observation could be made at around 30% of the CDF. In both UE- and network-centric schemes, a small step is observed near this value. It results from the difference in throughputs between the UEs with the "macro layer only" connection and all others. However, the network-centric solution additionally smoothens this difference due to its fairness-based optimization.

For the network-assisted scheme, this effect is not visible as the mechanism in question is non-bifurcated and the UEs have only one active connection at all times. However, another effect is noticed here: step-wise behavior may be observed in the upper part of the curve. This is due to the difference between the UEs with very good WiFi link qualities – various numbers of UEs associate with different APs causing variations in throughput if the number of connected UEs is low and the SNR of each user is high.

In Fig. 4, a more detailed performance comparison is conducted. To study performance dynamics, we evaluate the resource allocation approaches in three different scenarios: WiFi and macro LTE (left subplot), pico and macro LTE (central subplot), and finally a "fully heterogeneous" scenario integrating macro LTE, WiFi, and pico LTE. We were interested in varying the input parameters (by changing the $\beta$ coefficients, see Section III) of the network-selection schemes in order to investigate the resulting performance limits.

Naturally, increase in throughput causes degradation in the fairness levels, which is captured by the Jain's index on the vertical axis. Additionally, the limitations of the available trade-off opportunities are indicated with different colors. For instance, if one would attempt to increase fairness further, the coverage limitations will be met: no extra resources could be allocated to the UEs with poor connectivity, whereas the throughput of well-connected UEs will not give any additional benefits in fairness. On the other hand, if one would try to increase the throughput beyond the indicated limit, that is, to improve the throughput of well-connected users, the capacity limits will be met: no more resources will be allocated due to bandwidth constraints.

A similar trade-off is observed for the network-assisted scheme, when we vary the SNR-based association threshold for the LTE and WiFi small cells. However, we also learn that the network-centric strategy has a much wider balancing range, as well as higher resulting throughput and fairness performance. This practically means that the H-CRAN based

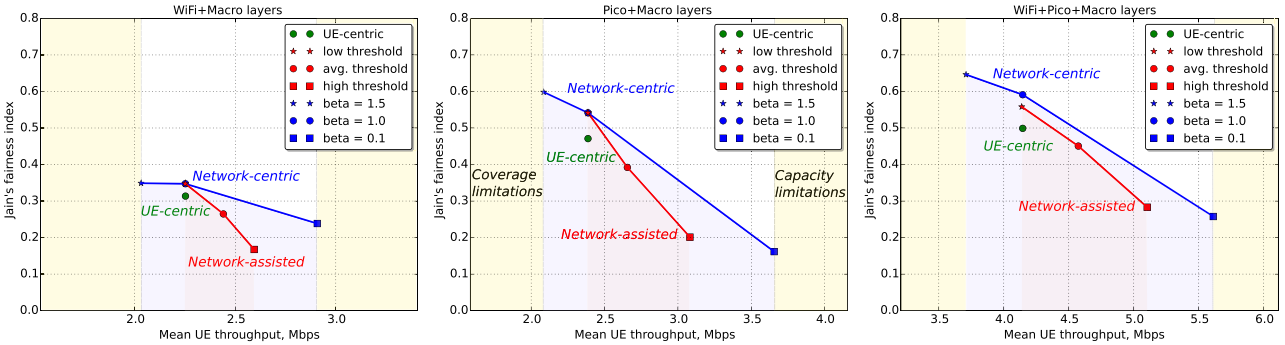



**FIGURE 4.** Performance and manageability comparison.

resource management performs better when more fairness is demanded or higher average throughputs are required. Apparently, the system has higher fairness and throughput values with more available RANs, but ''pico plus macro'' is more beneficial than ''WiFi plus macro'', even though the use of 20MHz WiFi bandwidth should have given better boost than 10MHz pico LTE channel. This is due to the fact that the initial coverage of the pico layer network is better (owing to higher transmission power and transmitter/receiver gains) together with the larger spectral efficiency of LTE technology.

## V. PROOF-OF-CONCEPT DEMO IMPLEMENTATION

To confirm practical feasibility of the discussed H-CRAN concepts and extend on the feasible options to integrate 5G-grade H-CRAN functionality into current mobile operator networks, our research group has recently completed a series of proof-of-concept prototypes. Those have resulted in implementation of an integrated testbed environment showcasing seamless integration of WiFi technologies into the existing 3GPP LTE network. This section summarizes our recent progress along these lines, addresses the major challenges faced, and offers important considerations on scaling the proposed technology solution for operator-wide 5G deployments.

Importantly, most of today's UE devices already support both cellular and WLAN radio interfaces. However, the strive to maximize energy efficiency has led the equipment manufacturers to impose a limitation in the mobile UE's operating system to have at most one radio interface active at a time. Currently, this situation changes as the dust around 5G wireless communications systems settles – 5G will be a highly synergistic integration of diverse radio-access techniques and solutions, rather than one killer technology. Consequently, the emerging demand for improved wireless connectivity has convinced major UE vendors to supply the developers with the adequate platform tools to access and effectively use all the available radio interfaces. A promising example of a research-friendly UE platform is Jolla phones running Sailfish OS. In our experiments, the devices under test proved to have very flexible and open architecture, augmenting modern hardware with capable developer tools to allow for intended modifications on system level and thus achieve the desired degrees of connectivity.

Further, to advance the vision of H-CRAN, we were seeking for the feasible options to integrate cellular and WLAN RATs. Ideally, integrating a WiFi access point on the network provider side would require access to an open cellular BS, which is hardly available in the research environment today. Hence, the need to efficiently aggregate both WiFi and cellular radios under the same operator network may be addressed by encapsulating the corresponding radio links into separate OpenVPN tunnels. These should be terminated on a virtual machine acting as the operator's aggregator node that might extend the functionality of the packet gateway or represent a dedicated entity within the BBU pool.

Our H-CRAN testbed design employs the emerging SDN architecture and, particularly, the OpenFlow protocol to dynamically and efficiently manage UE connectivity [14]. All of the radio links available to a particular UE are treated as connected to a single forwarding plane on the phone, which is emulated by running the Open vSwitch software. The actual forwarding decisions are made by the controller software basing on a set of active measurements of the radio link conditions; they are implemented in the switch following the OpenFlow rules. In our test setup, the controller is running on the UE itself, but the proposed solution is more general and the UE configuration can easily be outsourced to the network provider running a single controller for a set of all served users. However, the latter option would naturally impose additional requirements on the control channel availability.

More specifically, the anchor points for the VPN tunnels on the operator side are running in isolation inside the respective Linux containers, which are maintained by the Docker Engine. The virtual backhaul links from the containers are plugged into the Open vSwitch daemon emulating the operator's forwarding plane. By design, Open vSwitch assumes that all the links under virtual interfaces are able to handle Ethernet frames [14], but the cellular link is exposed to the system by an RmNet or RmNet USB driver. Therefore, adding radio links to the virtual forwarding plane requires respective modification of the vSwitch daemon. In addition,

OpenFlow protocols need to count the desired protocol headers or skip the lower-layer headers with a predefined offset. In our testbed architecture, we have introduced an extra layer of abstraction on top of the links by the GRE tap tunnels. We remind that the GRE tap tunnels emulate the common layer 2 network segment between the UE and the aggregator on the operator side. Our overall demonstrator setup is detailed in Fig. 5 (bottom part).

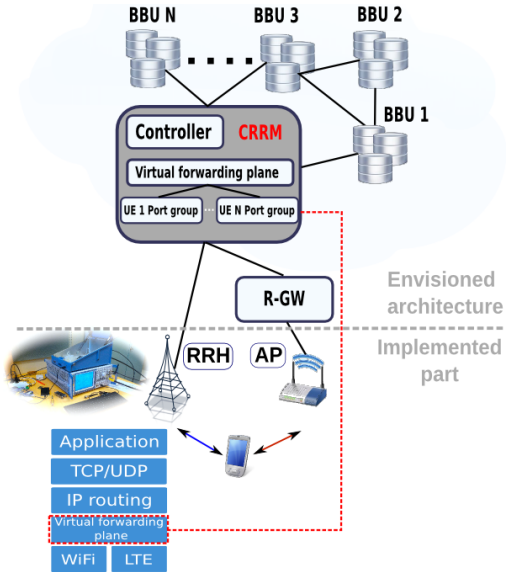



**FIGURE 5.** **Proposed demo prototype structure.**

In summary, the proposed combination of hardware virtualization mechanisms and operating system container-based virtualization methods enables quick and flexible deployment of various scenarios of interest within the context of emerging H-CRAN ecosystem. All of the necessary components are coupled in a highly configurable and integrative manner, which allows for our design to be easily reproduced in practical deployments. Conveniently, our proposed testbed design comprehensively supports the emerging vision of 5G network operator very recently offered in [15]. In particular, Fig. 5 outlines a characteristic deployment option for scalable integration of the discussed concepts, which have been explored with our testbed design, into the 5G-grade mobile operator's network.

## VI. CONCLUSION

In this article, we considered the concept of H-CRAN, which has recently emerged as cost-efficient solution to improve on the available cooperative gains in HetNets. We concentrated our attention on the problem of coordinated radio resource management in 5G-grade H-CRANs by providing a comprehensive methodology for their real-time performance optimization. The highlights of our solution are as follows.

- Our proposed resource optimization and control methodology is suitable for real-time resource allocations across multiple RANs in H-CRAN environment. It allows for flexible and adjustable balance between two major metrics – the overall H-CRAN system throughput and fairness of resulting allocations.
- The framework can be extended to address other important metrics of interest, including prioritization of users and RANs, non-bifurcated traffic solutions, finite traffic demands, etc.
- The optimization problem is of linear programming type allowing for real-time resource optimization in large-scale dense H-CRANs.
- Our performance comparison of resource allocation strategies is characteristic of realistic ANDSF, anchor-booster, and H-CRAN multi-radio integration mechanisms. H-CRAN integration is shown to be most beneficial in terms of throughput, fairness, and flexibility across a number of representative operator deployments.
- Basing on our review of technology implementation options and respective 3GPP standardization, the envisioned H-CRAN system architecture has been outlined together with a prototype implementation for the CRRM unit.

## ACKNOWLEDGMENT

The authors are grateful to Intel Finland, Rohde & Schwarz, and Jolla Oy for providing equipment necessary for the proof-of-concept implementation.

## REFERENCES

[1] J. G. Andrews *et al.*, "What will 5G be?" *IEEE J. Sel. Areas Commun.*, vol. 32, no. 6, pp. 1065–1082, Jul. 2014.

[2] Y. Yang and T. Q. S. Quek, "Optimal subsidies for shared small cell networks—A social network perspective," *IEEE J. Sel. Topics Signal Process.*, vol. 8, no. 4, pp. 690–702, Aug. 2014.

[3] O. Galinina *et al.*, "Capturing spatial randomness of heterogeneous cellular/WLAN deployments with dynamic traffic," *IEEE J. Sel. Areas Commun.*, vol. 32, no. 6, pp. 1083–1099, Jun. 2014.

[4] B. Bangerter, S. Talwar, R. Arefi, and K. Stewart, "Networks and devices for the 5G era," *IEEE Commun. Mag.*, vol. 52, no. 2, pp. 90–96, Feb. 2014.

[5] S.-H. Park, O. Simeone, O. Sahin, and S. Shamai, "Fronthaul compression for cloud radio access networks: Signal processing advances inspired by network information theory," *IEEE Signal Process. Mag.*, vol. 31, no. 6, pp. 69–79, Nov. 2014.

[6] J. Li, M. Peng, A. Cheng, Y. Yu, and C. Wang, "Resource allocation optimization for delay-sensitive traffic in fronthaul constrained cloud radio access networks," *IEEE Syst. J.*. [Online]. Available: http://arxiv.org/abs/1410.7867

[7] Y. Yang, T. Q. S. Quek, and L. Duan, "Backhaul-constrained small cell networks: Refunding and QoS provisioning," *IEEE Trans. Wireless Commun.*, vol. 13, no. 9, pp. 5148–5161, Sep. 2014.

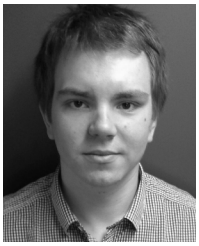
[8] Y. Zhou and W. Yu, "Optimized backhaul compression for uplink cloud radio access network," *IEEE J. Sel. Areas Commun.*, vol. 32, no. 6, pp. 1295–1307, Jun. 2014.

[9] S. Yan, W. Wang, Z. Zhao, and A. Ahmed, "Investigation of cell association techniques in uplink cloud radio access networks," *Trans. Emerg. Telecommun. Technol.*, 2014. [Online]. Available: http://onlinelibrary.wiley.com/doi/10.1002/ett.2894/abstract

[10] M. Peng, S. Yan, and H. V. Poor, "Ergodic capacity analysis of remote radio head associations in cloud radio access networks," *IEEE Wireless Commun. Lett.*, vol. 3, no. 4, pp. 365–368, Aug. 2014.

[11] Z. Ding and H. V. Poor, "The use of spatially random base stations in cloud radio access networks," *IEEE Signal Process. Lett.*, vol. 20, no. 11, pp. 1138–1141, Nov. 2013.

[12] A. Liu and V. Lau, "Joint power and antenna selection optimization in large cloud radio access networks," *IEEE Trans. Signal Process.*, vol. 62, no. 5, pp. 1319–1328, Mar. 2014.
[13] M. Peng, Y. Li, J. Jiang, J. Li, and C. Wang, "Heterogeneous cloud radio access networks: A new perspective for enhancing spectral and energy efficiencies," *IEEE Wireless Commun.*, vol. 21, no. 6, pp. 126–135, Dec. 2014.
[14] Open Networking Foundation. (Apr. 2012). *Software-Defined Networking: The New Norm for Networks*. [Online]. Available: https://www.opennetworking.org/images/stories/downloads/sdn-resources/white-papers/wp-sdn-newnorm.pdf
[15] P. Agyapong, M. Iwamura, D. Staehle, W. Kiess, and A. Benjebbour, "Design considerations for a 5G network architecture," *IEEE Commun. Mag.*, vol. 52, no. 11, pp. 65–75, Nov. 2014.



**MIKHAIL GERASIMENKO** received the Specialist degree from the Saint-Petersburg University of Telecommunications, in 2011, and the M.Sc. degree from the Tampere University of Technology, in 2013. He started his academic career in 2011. He is currently a Researcher with the Department of Electronics and Communications Engineering, Tampere University of Technology. He has co-authored multiple scientific journal and conference publications, and holds several patents. His main subjects of interest are wireless communications, machine-type communications, and heterogeneous networks. He acted as a Reviewer and participated in educational activities.

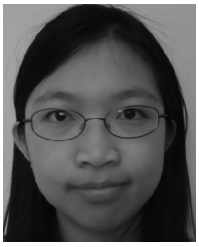
**DMITRI MOLTCHANOV** received the M.Sc. and Cand.Sc. degrees from the Saint-Petersburg State University of Telecommunications, Russia, in 2000 and 2002, respectively, and the Ph.D. degree from the Tampere University of Technology, Finland, in 2006. He is currently a Senior Research Scientist with the Department of Electronics and Communications Engineering, Tampere University of Technology. He has authored over 50 publications. His research interests include performance evaluation and optimization issues in wired and wireless Internet Protocol networks, Internet traffic dynamics, quality of user experience of real-time applications, and traffic localization in peer-to-peer networks. He serves as a TPC Member of a number of international conferences.

**ROMAN FLOREA** is currently a Research Assistant with the Department of Electronics and Communications Engineering, Tampere University of Technology, Finland. He holds a number of professional certifications in networking, and has wide experience in server operations for high load and high availability. His research interests include routing and switching in Internet Protocol networks, network operations, and the Internet coordination.

**SERGEY ANDREEV** received the Specialist degree in information security and the Cand.Sc. degree in wireless communications from the Saint-Petersburg State University of Aerospace Instrumentation, Saint Petersburg, Russia, in 2006 and 2009, respectively, and the Ph.D. degree in technology from the Tampere University of Technology, Tampere, Finland, in 2012. He is currently a Senior Research Scientist with the Department of Electronics and Communications Engineering, Tampere University of Technology. He has co-authored over 90 published research works. His research interests include wireless communications, energy efficiency, heterogeneous networking, cooperative communications, and machine-to-machine applications.

**YEVGENI KOUCHERYAVY** received the Ph.D. degree from the Tampere University of Technology (TUT), Finland, in 2004. He is currently a Full Professor and the Lab Director of the Department of Electronics and Communications Engineering with TUT. He has authored numerous publications in advanced wired and wireless networking and communications. His current research interests include various aspects of heterogeneous wireless communication networks and systems, the Internet of Things and its standardization, and nanocommunications. He is an Associate Technical Editor of the *IEEE Communications Magazine*, and an Editor of the IEEE COMMUNICATIONS SURVEYS AND TUTORIALS.

**NAGEEN HIMAYAT** received the B.S.E.E. degree from Rice University, in 1989, and the Ph.D. degree in electrical engineering from the University of Pennsylvania, in 1994. She has over 15 years of research and development experience in the telecom industry. She is currently a Senior Research Scientist with Intel Corporation, where she performs research on broadband wireless systems, including heterogeneous networks, cross-layer radio resource management, multi-antenna techniques, and optimizations for machine-to-machine communications.



**SHU-PING YEH** received the B.S. degree from National Taiwan University, in 2003, and the M.S. and Ph.D. degrees from Stanford University, in 2005 and 2010, respectively, all in electrical engineering. She is currently a Research Scientist with the Wireless Communications Laboratory, Intel Corporation. Her current research focus includes interference mitigation in multitier networks utilizing multiantenna techniques, machine-to-machine communications, and interworking of multiple radio access technologies within a network.

**SHILPA TALWAR** received the Ph.D. degree in applied mathematics from Stanford University, in 1996, and the M.S. degree in electrical engineering. She held several senior technical positions in the wireless industry. She has over 15 years of experience in wireless. She is currently a Principal Engineer with the Wireless Communications Laboratory, Intel Corporation, where she is conducting research on mobile broadband technologies. She has authored numerous technical publications and patents.