Virtualized cognitive network architecture for 5G cellular networks

Virtualized Cognitive Network Architecture for 5G Cellular Networks

Hesham ElSawy, Hayssam Dahrouj, Tareq Y. Al-Naffouri, and Mohamed-Slim Alouini

***Abstract*—Cellular networks have preserved an application agnostic and base station (BS) centric architecture[1] for decades. Network functionalities (e.g., user association), are decided and performed regardless of the underlying application (e.g., automation, tactile Internet, online gaming, multimedia). Such ossified architecture imposes several hurdles against achieving the ambitious metrics of next generation cellular systems. This paper first highlights the features and drawbacks of such architectural ossification. Then, the paper proposes a virtualized and cognitive network architecture, wherein network functionalities are implemented via software instances in the cloud, and the underlying architecture can adapt to the application of interest as well as to changes in channels and traffic conditions. The adaptation is done in terms of the network topology by manipulating connectivities and steering traffic via different paths, so as to attain the applications' requirements and network design objectives. The paper presents cognitive strategies to implement some of the classical network functionalities, along with their related implementation challenges. The paper further presents a case study illustrating the performance improvement of the proposed architecture as compared to conventional cellular networks, both in terms of outage probability and handover rate.**

***Keywords*:- Cognitive cellular networks, 5G, heterogeneous networks, small-cells, virtualized RAN, cloud RAN.**

I. Introduction

The fifth generation (5G) cellular networks is expected to offer ubiquitous and global connectivity for everything (users, devices, sensors, machines) and support diverse types of applications with different operational constraints. Some of these applications are delay sensitive (e.g., automation, tactile Internet), others are bandwidth (BW) aggressive (e.g., online gaming, multimedia), and others are demanding in terms of the numbers of connections (e.g., smart cities). Hence, context awareness (i.e., awareness of the application requirements and real-time network related information such as: the network conditions, source and destination relative locations, interference levels, congestion bottlenecks) is crucial for 5G networks to effectively support these diverse applications. However, this is not the case for conventional cellular networks, which have historically preserved an application agnostic and base station (BS) centric architecture. Regardless of the underlying application, network conditions, and relative locations of the devices, the network functionalities (e.g., user association, resource allocation, data routing etc.) are performed in the same manner. This may lead to an inefficient resource utilization in the radio access network (RAN) and unnecessary delays in the core network. Hence, such a problem, denoted in this paper by *architectural ossification*[2] spans both the access and core networks. This paper focuses on the RAN problems and discusses the potential solutions.

Architectural ossification may also impede the evolution towards the ambitious metrics defined for the 5G networks, namely, the 1000-fold capacity increase with at least 100-fold leap in the peak data rate and 0.1x delay reduction [1]. In particular, network densification via small-cells and millimeter wavelength (mmW) communication, which are the main drivers for capacity and data rate improvement, require flexible network architecture. For instance, the migration to the mmW band may impose spatial blind spots to BSs' coverage, due to the significant effect of shadowing on mmW propagation, which requires redundant associations to increase the probability of LoS connection (e.g., serving one user by several BSs). Also, conventional association in dense small-cell environment imposes considerable handover signaling due to mobility, hence, new association schemes are required for handover signaling reduction. Also, conventional association in dense small-cell environment imposes considerable handover signaling due to mobility, hence, new association schemes are required for handover signaling reduction.

To obtain the desired 5G performance metrics, cognition and flexible architecture realized via context aware network functionalities becomes a necessity. This paper integrates recent advances in cellular networking and proposes unified virtualized and cognitive architecture wherein the control and data planes are decoupled under a cloud-RAN (CRAN) umbrella, so as to adapt the network functionalities to changes in channels and traffic conditions as well as to the underlying application. Decoupling control and data planes, which is proposed in [2], enables centralized software control for the network behavior, i.e., instead of using certain modules to perform network functions at each and every BS, a single script written in the cloud can control the entire network behavior. The forwarding plane, however, remains distributed according to the physical locations of the network entities (i.e., BSs, servers, gateways). This leads to a self-organizing network (SON), and provides generous flexibility both in terms of network expansion via BS deployment, and in terms of creating new services and applications.

It is worth noting that centralized network control in the cloud does not necessarily mean a centralized execution for network functionalities. As will be discussed in Section III, some network functions may be implemented in a distributed cognitive manner while the cloud only dictates the operational guidelines. For instance, a proximity application using device-to-device communication can opportunistically access the cel-

Hesham ElSawy, Tareq Y. Al-Naffouri, and Mohamed-Slim Alouini are with King Abdullah University for Science and Technology (KAUST). Tareq Y. Al-Naffouri is also with King Fahd University of Petroleum and Minerals. Hayssam Dahrouj is with Effat University.

[1]Network architecture, in this paper, defines how BSs are connected to each other and how the cellular network elements (e.g., BSs, user, relays, etc.) communicate.

[2]We borrow the terminology form the *Internet ossification problem* due to the analogy between the two cases.

lular channel, in which the interference constraints for the cellular users are defined by the cloud[3]. In such a *virtualized, hybrid centralized and distributed architecture*, there is no single rule to execute network functions. Instead, network functions are performed based on the network condition and underlying application. According to the traffic requirements and bottleneck location, the cognitive cloud would steer the traffic from one path to another, or replace a single hop congested BS link with a multi-hop non-congested link to offer adequate quality-of-service (QoS) and meet the application constraints. For instance, an emergency automation message targeting an actuator in the proximity area around the sensor does not have to go through the serving BS and core network. Instead, it can be directly conveyed to the actuator in multi-hop fashion. Further, the cloud may suppress co-channel interference by nearby devices and BSs during the automation message transmission, so as to ensure an error free delivery for the emergency message.

CRAN is also an enabler for proactive, instead of reactive, network control. The cloud can gather information (e.g., from social networks) and learn about users' interests, preferences, data usage, and mobility patterns. This information can be used for proactive resource allocation and data caching, which can substantially enhance the network performance [3]. Fig. 1 illustrates the proposed architecture and emphasizes the different supported services. The figure shows the different network connectivity types (i.e., multi-hop, device-to-device, control signaling decoupling), in which the control engine is located at the cloud. Context awareness at the cloud enables the network to respond differently to different applications and network conditions.

The remainder of the paper presents practical strategies of implementing cognitive network functionalities. Then it introduces the related implementation issues. The paper finally shows potential gains of the proposed virtualized cognitive architecture through illustrative simulations.

II. Network Functionalities

From the RAN perspective, network topology is defined by the connections between the network entities. Fig 2(a) shows the star-shaped topology enforced by conventional cellular networks. The star-topology emerges from the fact that all transmissions are routed through BSs, regardless of the underlying application and the required data flow, in a centralized manner to guarantee efficient interference management. The communication within the conventional cellular network is merely "*BS centric*" where cell boundaries determine the region served by each BS. The relative radio signal strength (RSS) between neighboring BSs is the common way to determine cell boundaries. This section revisits some of the classical network functionalities and shows how redefining such functionalities in a cognitive fashion yields an appreciable flexible network operation as shown in Fig 2(b).

[3]It is important to highlight that estimating the channel gains towards the receivers, which is infeasible in TV white space cognitive network due to their passive TV set receivers, may be feasible in cellular networks. This is because cellular networks are characterized by their active elements, which simultaneously transmit and receive data. Hence, the channel can be estimated during the transmission period.

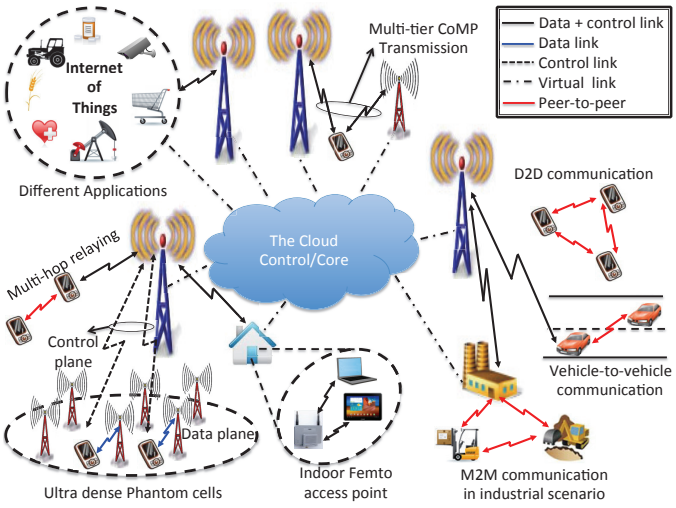
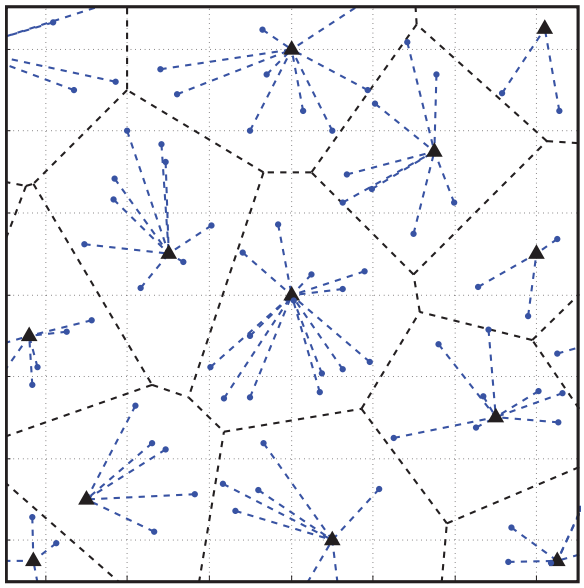


Fig. 1. Schematic digram for 5G cellular network. The cloud is the core engine of the network which monitors the traffic spatial and temporal variation as well as the traffic classes, based on the underlying application, and controls the network operation accordingly.

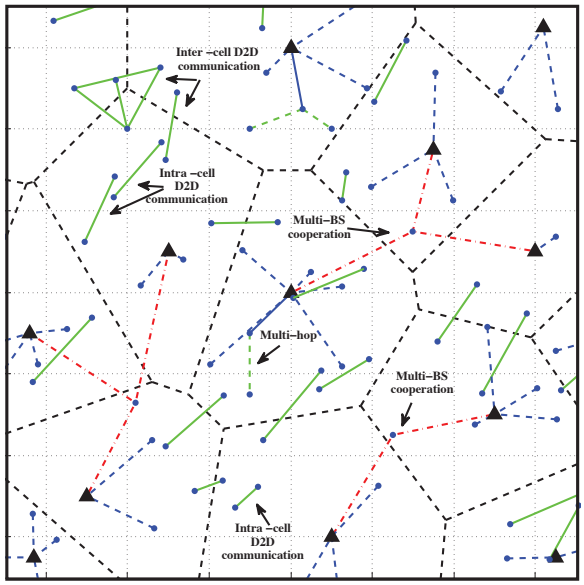
A. User Association

1) Conventional Operation: User association is the most basic, yet critical, network function in cellular networks, as it assigns users to BSs. User association controls the traffic and load served by each BS and each network tier. Association impacts many performance metrics such as coverage probability, rate, blockage, etc. Therefore, many attempts have been made to optimize BS association in cellular networks [4], [5]. The common message in these works is that RSS based association is not always efficient. For instance, in [4], the authors show that different users' association strategies are required to attain different design objectives (e.g., delay, rate, fairness) in a single tier cellular networks. In [5], it is shown that the inefficiency of RSS based association is more prominent in multi-tier networks due to the high transmit power disparity between BSs types (i.e., macro, micro, pico, and femto). In fact, the authors show that, with the proper manipulation of the association function, in a two tier cellular network, the minimum rate attained by the users can be increased by 4 times.

2) Proposed Operation: The virtualized network architecture not only renders the RSS rule obsolete and allows a flexible cell association, but also allows decoupled uplink, downlink, and control association. To illustrate how virtualization achieves such decoupling, consider a three-tier network in which a test user has a pico-BS as his closest BS, then a micro-BS farther than the pico-BS, and then a macro-BS farther than the micro-BSs. However, due to the downlink transmit power disparity, the macro-BS (pico-BS) provides the highest (lowest) downlink RSS. Instead of associating with the macro-BS only, which might be congested, the user can communicate in the downlink with the micro-BS for load balancing, and in the uplink with the pico-BS for transmit power reduction. To reduce the handovers caused by mobility, the user can receive control signaling from the macro-BS. Cognition, in



(a) BS centric topology (*conventional star-topology*) where each UE connects to its nearest BS



(b) Context aware topology in which the connections are established based on several aspects such as the relative distance between nodes, application, SINR, ...etc.

Fig. 2. Network topologies for the same locations of BSs and UEs in which the triangles represent the BSs and the dots represent users, black dotted lines represent cell boundaries, blue dotted lines represent single BS connectivity, red dotted lines represents multi-BS connectivity, and greed dotted lines represent peer-to-peer D2D connectivity.

this case, becomes important, since there is no single rule for association, as it depends on the underlying application and the network conditions. For example, if an application has tight rate constraint, an uplink connection to a less loaded, although much farther and may require higher uplink transmit power, BS may be more efficient than a congested nearby BS. Further, users' association has to adapt to the traffic and spatial distributions in order to attain the desired network objective and application requirements. As shown later in section IV, significant gains can be harvested from such a flexible association.

B. Device-to-Device Communication

1) Conventional operation: The conventional cellular infrastructure dictates that all traffic are communicated through the BSs, regardless of the users' relative locations (i.e., the required data flow). However, with the proliferation of proximity based services and social networking, nearby users may wish to establish connections and exchange data. Recent studies show that if nearby devices are allowed to bypass the cellular infrastructure and directly communicate in a peer-to-peer fashion, which is referred to as device-to-device (D2D) communication, many performance metrics can be improved [6]. D2D communication has the potential to offload traffic from congested cellular BSs and spatially reuses short-distance small-power peer-to-peer links. Recent studies [6] suggest that D2D communication can improve the cellular system throughput by 374%, the power efficiency by 100%, and the cell edge users' performance by 300%, when compared to conventional star-topology. It is worth mentioning that D2D communication is the key enabler for the cellular network to support machine-to-machine (M2M) communication and Internet of things. D2D communication enables massive number of machines to communicate together and connect to the Internet. Admitting all of the machines traffic to the cellular infrastructure overloads the network and results in congestion and blockage. Offloading such machine-type communication to the D2D mode, whenever possible, alleviates the congestions and enhances the overall network performance.

2) Proposed Operation: The proposed virtualized network architecture provides a flexible paradigm for enabling D2D connectivity. That is, on top of the decoupled uplink, downlink, and control signaling connectivity, there is an option to establish one or more of these connections in the D2D mode. For instance, consider two low power devices A and B such that A is located between B and a nearby BS. If A is closer to the BS than to B, the transmission from A to B can be established via the BS and the transmission for B to A can be established in the D2D, and hence, the transmit power is reduced. Such flexible communication paradigm enables fine tuned traffic control to reduce congestions, reduce power consumption, and increase the network efficiency. However, D2D communication increases the complexity of the admission process as it defines a new network function. That is, the admission process of each user includes a mode selection function to determine the mode of operation of each link, namely, D2D mode or cellular mode. The D2D communication also changes the conventional cellular star-topology to a hybrid topology (i.e., coexisting star and ad hoc topologies), in which interference management is the core challenge. In this case, cognition is important for interference management between D2D and cellular links. Note that the priority of the D2D and cellular links to use the spectrum is based on the application. For instance, as discussed earlier, a critical D2D automation message would be the primary spectrum user and the cellular links would be the secondary

spectrum users. On the other hand, a file transfer via D2D mode would be the secondary spectrum user and the cellular links would be the primary users. Also, cognitive coordination between different D2D links for spectrum access is important to maintain an acceptable QoS.

C. Multicell Coordination

1) Conventional Operation: Conventional network architecture enforces one user to one BS association strategy. In this case, employing aggressive frequency reuse schemes, due to the scarcity of the wireless spectrum, imposes high inter-cell interference. Such interference is a main performance limiting parameter for cellular networks, specially for cell edge users. Multi-cell cooperation, via Coordinated Multi-Point transmission (CoMP), is employed to reduce intercell-interference and improve the signal-to-interference-plus-noise-ratio (SINR) statistics. That is, multiple BSs cooperate (e.g., via beamforming) to simultaneously serve multiple common users. Such cooperation boosts the system capacity by reducing inter-cell interference and improving the SINR statistics[4].

2) Proposed Operation: The virtualized network architecture enables multiple BS association for each of the uplink, downlink, and control signaling for each user. CoMP changes the network topology to mitigate inter-cell interference and enhance cell edge performance as shown in Fig 2(b). However, this may increase the signaling overhead between BSs. In this case, per user cognitive enabled/disabled CoMP scheme can be implemented. This leads to a flexible CoMP operation that accounts to the user SINR, the application requirement, and the state of the surrounding BSs (e.g., congested or not). Note that, in a CRAN environment, BSs act as virtual antenna system for the cloud, which transforms CoMP scheduling to a multiuser (MU) MIMO scheme.

D. Multi-Hop Relaying

1) Conventional operation: Earlier, we discuss the D2D mode of communication as an alternative to the cellular link. A potential application for the D2D mode is multi-hop relaying. That is, a cellular link can be replaced by several consecutive D2D links. Multi-hop relaying is essential with the drastic increase in the population of devices. There could be situations where massive number of machines simultaneously need to connect to the cellular infrastructure. One example is traffic jams where all smart vehicles need to connect to the Internet to send information or receive updates about the traffic conditions. Another instance may occur in over crowded places like stadiums where a huge population of users want to simultaneously connect to the network. In these cases, the direct link connectivity imposed by the conventional cellular infrastructure results in high blockage probability and degraded user experience. Replacing the single direct link per channel per cell to multiple short-distance low-power links that can be reused several times within the same cell area alleviate such congestion. It is shown in [7] that the appropriate design of multi-hop relaying can increase the minimum achievable rate of users by up to 100 fold.

2) Proposed Operation: The virtualized network architecture allows multi-hop relaying in each of the decoupled links. This can be exploited to extend BS coverage, where users in coverage holes can relay information to the BS via another covered user, or in high connectivity demand periods. An important application for multi-hopping is proximity-based services with delay constraints. In some cases, the BSs are located hundreds of kilometers away from the nearest serving gateway (SGW) [8]. Hence, routing information in the conventional way via the core network (i.e., SGW) may encounter unnecessarily high delays. Therefore, multi-hop relaying is a key solution to satisfy the delay constraint for such applications. Multi-hop relaying defines a new network function that selects whether to multi-hop or directly send to the BS. Hence, multi-hopping changes the star-topology into an extended star-topology, which imposes an inherent routing problem. The cloud should be able to detect network congestion, in terms of users' population, and enable multi-hopping to maintain an acceptable blocking probability. As discussed earlier, the priority for the channel access between cellular and multi-hop D2D links highly depends on the underlying application.

III. Implementation Issues

The massive number of network elements (i.e., BSs, users, machines, etc.) makes a centralized instantaneous optimization for the network functions, even with super computing agents in the cloud, infeasible. That is, it is infeasible to select serving BS, assign powers, allocate channels, and choose the mode of operation for every and each network element within the cellular network. In this section, we discuss some trends to compromise between complexity and performance in prospective cognitive CRAN networks as well as the limitations for CRAN operation.

A. Feasibility and Complexity Tradeoffs

For feasible and efficient network operation, we seek tradeoffs between complexity, signaling, and performance. We advocate to split the network functions into two main categories, namely, instantaneously and statistically optimized functions. While instantaneous optimization guarantees best performance at any time instant, statistical optimization provides optimal averaged performance on long-time scale to reduce signaling and processing overheads. Specifically, instead of requiring instantaneous information, which is difficult to obtain and communicate, statistical network parameters can be exploited to guarantee an average optimal performance. For instance, in [9], instead of optimizing the transmit power based on the instantaneous channel gains, a simple power policy, which is optimal to the channel gain distribution, is developed. It is worth mentioning that many attempts have been made to either statistically or instantaneously optimize network functions. However, to the best of the authors' knowledge, merging statistical and instantaneous optimization to balance performance, complexity, and signaling overhead has been ignored.

[4]Several cellular operators have already launched CoMP trials, also denoted as elastic cell, and demonstrated the performance improvement.

In addition to the statistical optimization, cognitive and distributed control for some network functions can be exploited to reduce complexity. In this case, the network elements can choose their instantaneous operating parameters (e.g., power level, channel, and mode of operation) in a cognitive way to maximize the network objective subject to the enforced operator policies. For example, based on the relative D2D and cellular link distances, users can distributively select their mode of operation. Then, on one hand, D2D users can operate in a distributed and cognitive manner. On the other hand, antenna selection, channel assignment, and power control for the cellular mode users can be centrally controlled by the cloud. The cloud enforces a mode selection and operation policy rather than allocating channels and power levels for each D2D user. The policy typically dictates the conditions at which users select their operation mode and/or enforces interference limits for D2D users on cellular users. One approach to supervise the distributed control of network functions is through bias factors, which encourage/discourage users to take certain actions, as discussed in the next subsection.

B. Biased Network Functionalities

The association strategy can be manipulated via tunable bias factors that artificially encourage users to associate to a certain tier for each of the uplink, downlink, and control signaling. These bias factors can expand the coverage regions of small-cells to proportionally balance the load served by each network tier. Also, a control association bias factor towards higher network tiers (i.e., micro and macro), which is proportional to the user mobility, can used to reduce handover signaling. It is worth mentioning that CoMP can be considered as multiple BSs association which can be also controlled via bias factors. In this case, each user is responsible to report the set of candidate serving BSs to the cloud, which then manages the resource allocation for that user within the selected BSs. Tuning the bias factor that encourages/discourages cooperation controls the extent to which cooperation is enabled in the network, to tune the tradeoff between the SINR performance of CoMP and the associated backhaul traffic [10].

Similar to the user association, the D2D communication and multi-hop relaying can also be manipulated via bias factors that encourage/discourage transmitters to select single/multi-hop D2D communication. Setting the bias factor to zero enforces direct BS communication and setting the bias factor to a high value enforces D2D communication and multi-hopping. Hence, the bias factor can be tuned according to the traffic load and population to control the number of hops, delay, and hop distance.

Application dependent bias factors can be exploited to enforce general operator policies (i.e., objective and constraints) and guide the cognitive behavior of each application. The bias factors are calculated and adapted in the cloud, according to the traffic's spatial and temporal variation, and then dictated to the network elements.

C. Limitations of the Cloud Operation

As shown in Fig. 1, the cloud acts as an engine that adapts the proposed virtualized architecture to the type of applications and network conditions. A written script running at the cloud controls the network behavior. As discussed earlier, some network functions may be centrally executed in the cloud and others may be executed in a distributed manner (i.e., by BSs or devices). Note that, information about devices, BSs, and network conditions are required at the cloud to dictate the guidelines for distributed network functions and to execute centralized network functions. This information is subsequently communicated to the cloud by network entities via backhaul links, e.g., fiber optical cables, wireless backhaul links.

While fiber connections may be suited for cells of medium to large size (micro and macro-cells) [11], their deployment cost becomes a major problem for hundreds or thousands of small-cells BSs, as is the case in dense 5G networks. Furthermore, even when available in abundance in urban areas, fiber optical cables may not be found at the exact location where a small-cell BS exists. Wireless backhaul links, on the other side, are more suitable to support small-cell networks, as they are easy to plan and deploy when compared to fiber optics [12]. However, given the scarcity of the available sub-6 GHz licensed spectrum band, investment in higher frequency ranges is needed, which itself leads to a smaller coverage and needs fine tuning (i.e. strong LoS path) between the cloud and the served entities. Most importantly, a major problem in wireless backhaul design is the latency issue, as retransmissions are often required due to unsuccessful reception [13]. Additional latency is also added due to long round trip information routing to the data centers (i.e., cloud physical locations). In this case, routing delays can be reduced via multi-cloud location planning [8].

The above factors (i.e. cost, latency, coverage) often limit the operation of the cloud, and require intelligently jointly designing both the core (clouds to BSs) and RAN (BSs to users). Besides multi-cloud cooperation and coordination, synchronization, and joint provisioning of resources between the backhaul links and the RAN are promising future research directions, so as to ensure the feasibility of the virtualized cognitive architecture.

IV. Numerical Results

In this section, we show the potential gain of decoupling the uplink, downlink, and control associations. While uplink and downlink decoupling improves the SINR statistics, decoupling control association reduces signaling overhead [2]. Fig. 3 shows the gain in terms of outage probability if uplink and downlink associations are decoupled. We consider two-tier cellular network with channel inversion power control, in which all users maintain an average power of ρ at their serving BSs. The figure shows that if users associate in the uplink based on the uplink RSS, rather than the downlink RSS, the SINR outage probability can be reduced by 30%. Note that the outage probability for both the coupled and decoupled association increases with the intensity of small-cells due to the increased number of interferers and the fixed received power at the test BS (i.e., ρ).

Consider the two tier network shown in Fig. 4 in which a user follow the trajectory highlighted in black. In conventional

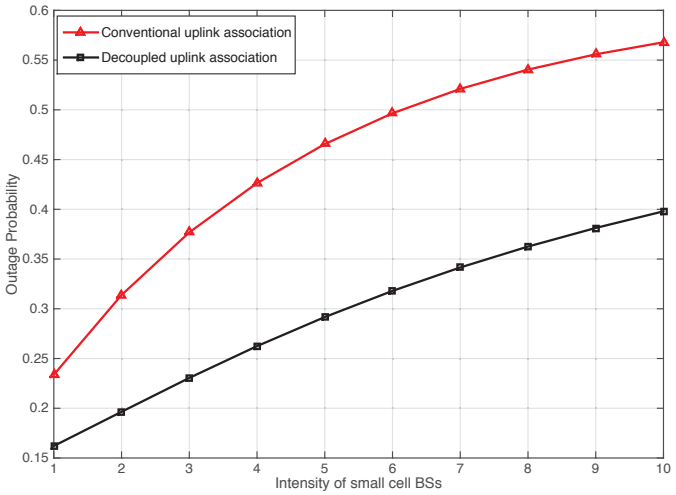


Fig. 3. Outage probability vs small-cell intensity for virtualized and conventional cellular architecture, the curves are obtained via stochastic geometry analysis following the footsteps of [14].

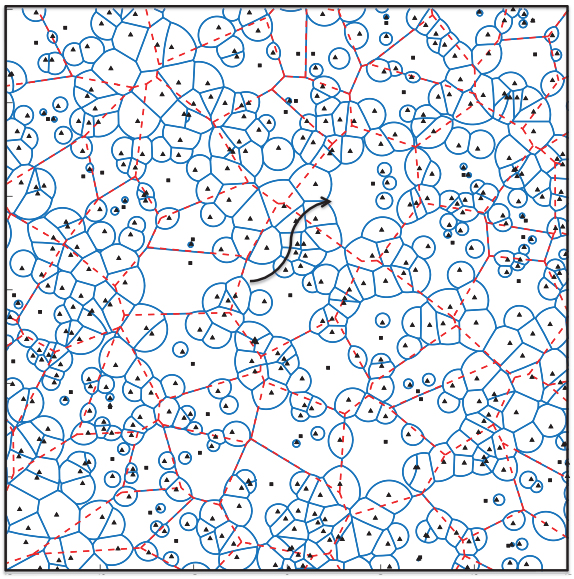


Fig. 4. A two tier cellular networks with macro-BS (squares), small-cells (triangles), and a user's trajectory (highlighted in black). The figure shows the handover boundaries (in blue) for the conventional cellular network architecture and handover boundaries (in dotted red) for the virtualized cellular network architecture.

network architecture, the test user performs a complete handover (i.e., dissociate from the serving BS, associate to the target BSs, and inform the core network for data flow switching) with every crossing over a cell boundary. Decoupling the control and data allows the macro-BS to act as a mobility anchor providing control signaling while the small-cells only provide data packets (which is referred to as *lean carrier* in the literature) [2]. In this case, complete handovers take place in transitions between macro-cells only (i.e., the red boundary shown in Fig. 4). In contrast, only *virtual handovers* (i.e., only data packet switching between BSs) take place in transitions involving small-cells. In a virtual handover, the macro-cell

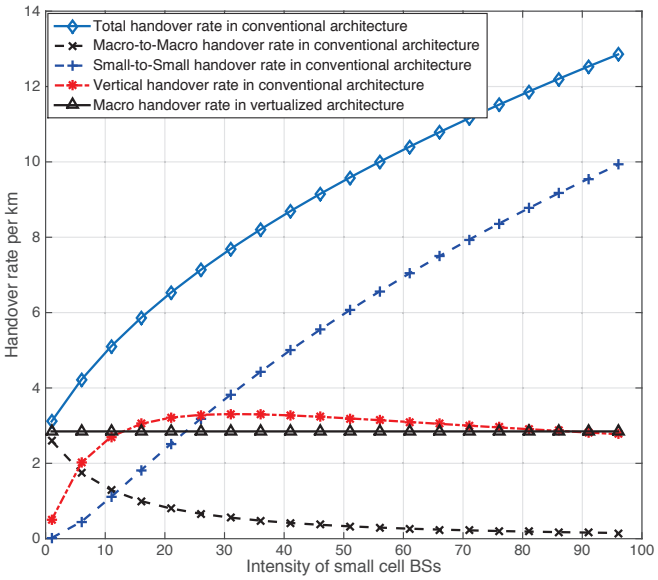


Fig. 5. The handover rate per km for conventional and virtualized network architecture, the curve is plotted via stochastic geometry analysis following the footsteps of [15] .

acts as handover anchor and the core network is not informed (i.e., the MME and SGW), which reduces the handover delay and core network signaling. Further, macro-BSs, have large coverage areas which decreases the complete handover rate. In fact, complete handover rate becomes independent of the small-cells intensity. To show the amount of handover reduction by virtualized network architecture, we plot Fig. 5. The figure shows that the complete handover rate linearly increase in the small-cell intensity. The handover rate is dominated by horizontal small-cell to small-cell handover. On the other hand, macro-to-macro handover reduces as the macro boundaries are populated with small-cells. For the virtualized network architecture, the complete handover rate is constant. Hence, the visualization gain increases with the intensity of small-cells. Note that reducing the handover rate can be directly translated to reduced delay and increased throughput. Hence, higher capacity gains can be harvested from network densification. It is worth noting that cognition may play an important rule in selecting the BS providing the control information. This is because, in a dense small-cell deployment, macro-BSs may have insufficient BW to provide the control signaling for all users. Hence, the control signaling for high mobility users only is handled by macro-cells. Control signaling for lower mobility profiles are handled by smaller BSs (i.e., micro and pico).

V. Conclusion

Architectural ossification for cellular networks is a limiting performance obstacle. Cognitive and flexible network operation via data/control plane decoupling and CRAN is an appealing trend to overcome architectural ossification problem. In this case, the cellular networks evolve from deterministic infrastructure to flexible, cognitive, and context aware architecture. Such an evolution is expected to improve the network

performance and derive the foreseen 5G performance gains. To this end, this paper discusses implementation of different network functions and sheds light on the tradeoffs between complexity, signaling and performance. The paper proposes distributed network control scheme in which the cloud may allow a guided distributed execution for network function, in which the guidelines are dictated by the cloud according to the application, operator policies, traffic and network conditions. The paper also shows that bias factor based guidance for network functions can be used to achieve a certain design objectives. Finally, the paper presents a case study to show the performance gain from decoupling uplink, downlink, and control association in multi-tier cellular network.

VI. Acknowlagement

The work of M. -S. Alouini was supported by the Qatar National Research Fund (a member of Qatar Foundation) under NPRP Grant NPRP 5-250-2-087. The work of T. Y. Al-Naffouri was a supported by KAUST project no. EE002355 at the Research Institute, King Fahd University of Petroleum and Minerals. The statements made herein are solely the responsibility of the authors

References

[1] J. G. Andrews, S. Buzzi, W. Choi, S. Hanly, A. Lozano, A. C. . K. Soong, and J. C. Zhang, "What will 5G be?" *To appear in IEEE J. Sel. Areas Commun. Issue on 5G Wireless Commun. Systems*, Sep. 2014, on archive: http://arxiv.org/pdf/1405.2957v1.pdf.

[2] Y. Kishiyama, A. Benjebbour, T. Nakamura, and H. Ishii, "Future steps of LTE-A: Evolution toward integration of local area and wide area systems," *IEEE Wireless Commun. Magazine*, vol. 20, no. 1, pp. 12–18, Feb. 2013.

[3] E. Bastug, M. Bennis, and M. Debbah, "Living on the edge: The role of proactive caching in 5g wireless networks," *IEEE Communications Magazine*, vol. 52, no. 8, pp. 82–89, Aug 2014.

[4] K. Hongseok, G. d. Veciana, Y. Xiangying, and M. Venkatachalam, "Distributed α-optimal user association and cell load balancing in wireless networks," *IEEE/ACM Trans. on Networking*, vol. 20, no. 1, pp. 177–190, Feb. 2012.

[5] H.-S. Jo, Y. J. Sang, P. Xia, and J. Andrews, "Heterogeneous cellular networks with flexible cell association: A comprehensive downlink SINR analysis," *IEEE Trans. Wireless Commun.*, vol. 11, no. 10, pp. 3484–3495, October 2012.

[6] A. Asadi, Q. Wang, and V. Mancuso, "A survey on device-to-device communication in cellular networks," *IEEE Commun. Surveys and Tutorials*, pp. 1–19, 2014.

[7] H. Tabrizi, G. Farhadi, and J. M. Cioffi, "A framework for spatial reuse in dense wireless areas," in *IEEE Global Telecommun. Conf. (Globecom 2013)*, Atlanta, GA, USA, Dec. 2013.

[8] A. Blenk, A. Basta, W. Kellerer, T. Zinner, F. Wamser, and P. T.-Gia, "Applying NFV and SDN to LTE mobile core gateways; the functions placement problem," in *ACM Special Interest Group on Data Communication (SIGCOMM'14), 4th Workshop on All Things Cellular: Operations, Applications and Challenges 2014*, Chicago, USA, Aug. 2014.

[9] C. Yu, O. Tirkkonen, K. Doppler, and C. Ribeiro, "On the performance of device-to-device underlay communication with simple power control," in *Proc. of IEEE 69th Vehicular Technology Conference (VTC Spring 2009)*, Barcelona, Spain, 2009.

[10] A. Sakr, H. ElSawy, and E. Hossain, "Location-aware coordinated multipoint transmission in OFDMA network," in *IEEE Intern. Commun. Conf. (ICC 2014)*, Sydney, Australia, Jun. 2014.

[11] Y. Shi, J. Zhang, and K. B. Letaief, "Group sparse beamforming for green cloud-RAN," *IEEE Trans. Wireless Commun.*, vol. 13, no. 5, pp. 2809–2823, May 2014.

[12] H. Dahrouj, W. Yu, T. Tang, J. Chow, and R. Selea, "Coordinated scheduling for wireless backhaul networks with soft frequency reuse," in *European Signal Processing Conference (EUSIPCO)*, Sep. 2013, pp. 174–177.

[13] D. Chen, T. Quek, and M. Kountouris, "Wireless backhaul in small cell networks: Modelling and analysis," in *in IEEE 79th Vehicular Technology Conference (VTC Spring)*, May 2014, pp. 1–6.

[14] H. ElSawy and E. Hossain, "On stochastic geometry modeling of cellular uplink transmission with truncated channel inversion power control," *IEEE Trans. Wireless Commun.*, vol. 13, no. 8, pp. 4454–4469, Aug. 2014.

[15] J. Bao and B. Liang, "Stochastic geometric analysis of user mobility in heterogeneous wireless networks," *To appear in IEEE J. Sel. Areas Commun. on Recent Advances in Heterogeneous Cellular Networks*, Apr. 2015, http://www.comm.utoronto.ca/ liang/research.html.