A Service-tailored TDD Cell-Less Architecture

Vincenzo Sciancalepore*, Konstantinos Samdanis*, Rudraksh Shrivastava*‡, Adlen Ksentini†, Xavier Costa-Perez*
* NEC Europe Ltd., Germany † Eurecom, Sophia Antipolis ‡ Dept. of Electronics, University of York, UK
{vincenzo.sciancalepore,samdanis,rudraksh.shrivastava,xavier.costa}@neclab.eu, {adlen.ksentini}@eurecom.fr

***Abstract*—The emerging 5G systems are envisioned to support higher data volumes and a plethora of different services with diverse QoS demands. To accommodate such service requirements, a cost efficient and flexible network architecture considering different service types is desired. The adoption of C-RAN can reduce infrastructure costs especially for dense deployments while at the same time centralize and hence optimize certain operations related with the control and data plane of the associated cells. This paper investigates such C-RAN approach in the context of TDD networks enabling a cell-less experience for users residing within overlapping areas. In particular, users are allowed to utilize selected sub-frames from different cells forming, in this way, a customized cell-less frame in a flexible manner. A queueing model and analysis is provided for optimizing power control and delay targets. A simulation study shows that our cell-less proposal significantly advances the state of the art both in terms of application and system performance.**

I. Introduction

As mobile communications enter a new area with more diverse applications, e.g., Internet of Things (IoT), beyond the traditional human communications, operators are facing increasing traffic volumes, which force them towards enhancing their network infrastructure dense deployments of support flexibly customized networking [1]. However, new Radio Access Network (RAN) elements that offer higher capacities in particular areas hugely increases capital and operational expenditures without fully supporting the requirements for a service-tailored flexibility. The use of remote radio heads and centralized base-band processing in combination with cloud infrastructures, i.e. Cloud-RAN (C-RAN), can significantly reduce costs while providing service-tailored flexibility by centralizing control-plane functions and offering efficient interference control as well as resource allocation.

In addition, C-RAN enables a scalable support of LTE-A features, such as Cooperative Multi-Point (CoMP), and emerging communication approaches, such as dual connectivity. Effectively, dual connectivity gives rise to a cell-less architecture, in where the control and data planes are not necessarily associated with a single cell but spread across a number of overlapping cells. This paper explores such cell-less paradigm considering a Time-Division-Duplex (TDD) system similar to the 3GPP enhanced Interference Management and Traffic Adaptation (eIMTA) [2], which allows neighboring cells to adopt an independent dynamic uplink (UL)/downlink (DL) re-configuration while reflecting evolving traffic demands as well as considering interference mitigation. However, differently from the distributed operation of eIMTA, our proposal abstracts the control-plane into a C-RAN baseband unit that is responsible for performing power control and resource allocation selecting also the UL/DL ratio on behalf of the associated cells.

This project has been partially funded from the European Unions Horizon 2020 research and innovation programme under the grant agreement No. 671584 5GNORMA and from the European Commmission ITN FP7 Marie Curie project CROSSFIRE (MITN 317126).

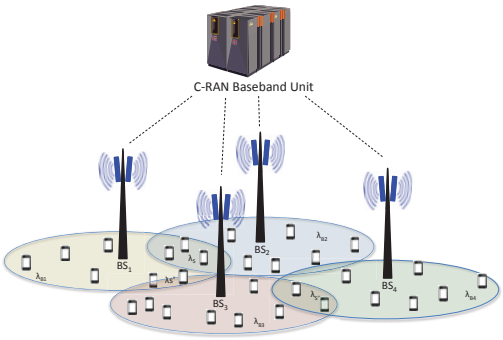


Fig. 1: C-RAN-based TDD Cell-less Architecture.

The initial idea of separating the control from the data plane is introduced with the phantom cell idea in [3] boosting the User Equipments (UEs) Quality of Service (QoS) through the physical resources of small cells while maintaining the control plane signal at macro-cells higher coverage. A step further considering more generic control-plane centralization based on Software Defined Networking (SDN) is documented in [4], focusing on interference control, mobility, and resource allocation. An equivalent resource abstraction concept is introduced by V-Cell [5], offering heterogeneous access resources as a pool to UEs that perceive connectivity as a single logical cell.

Our approach is based on the virtual cell concept in TD-LTE systems [6] [7], carrying out a data-plane abstraction wherein users opportunistically select subframes from overlapping cells. UEs residing in overlapping cell regions perceive such a resource allocation as a single logical cell, which can reflect best the required UL/DL demands and Quality of Experience (QoE), as analyzed in [8]. A higher degree of flexibility in resource allocation is realized as mobile operators can offer diverse UL/DL ratios based on application type requirements for particular geographical regions.

This paper extends such a virtual cell approach on a C-RAN architecture wherein the centralized entity abstracts the control-plane operations, including interference management and resource allocation. The novel contributions can be summarized as follows: (i) a cell-less architecture considering a C-RAN paradigm, (ii) a queuing model that helps to analytically evaluate the performance of the proposed cell-less architecture, (iii) a powerful C-RAN-based framework to dynamically

adjust the transmission power levels of the controlled base stations and (iv) an empirical performance study comparing load balancing considering target performance measure for video applications.

The remainder is organized as follows. Section II elaborates the cell-less architecture. Section III describes the queueing model of the propose cell-less architecture, while Section VI analyzes the performance evaluation and results. Finally, Section V concludes the paper.

II. Cell-less architecture

The virtual cell concept defines the property of UEs to opportunistically utilize resources from multiple base stations to satisfy their traffic demands. This makes room to a cell-less architecture within TDD networks, as described in [6], [7]. Typically, individual cells follow a specific TDD configuration that can dynamically be adjusted considering an UL/DL ratio that best matches the traffic requirements. Conversely, UEs residing within overlapping regions may have a data plane associated with multiple cells. Such operation abstracts a customized TDD frame issued by a virtual cell for the set of served users. The benefits are evident in cases of pseudo congestion, i.e., when resources are available on the other transmission direction, while interference mitigation can be achieved via a power control scheme [9].

The cell-less concept is efficiently cast into the C-RAN architecture, as depicted in Fig. 1, wherein a centralized baseband unit manages the resource allocation, power control and the configuration of the UL/DL ratio offering a seamless user experience. Cell-less frames are available in the overlapping area between multiple base stations so that users may opportunistically use UL and DL resources of different base stations by following the C-RAN baseband unit guidelines.

The centralized user resource allocation and frame reconfiguration requires signaling mechanisms to synchronize UEs and align the transmission and reception modes, accordingly. Indeed, the synchronization of UL/DL direction is required to ensure that the data to and from the UE appears as a single stream. A C-RAN baseband unit may decide how to distribute the incoming traffic and instruct UEs to forward upstream traffic in order to achieve such a synchronization. Once cell-less frames are configured and resource allocation is completed, mechanisms to perform management and maintenance are necessary to face with dynamic traffic demands. The goal is to assess the current UL/DL traffic load while deciding on-line whether a reconfiguration might improve the overall system performance. In addition, the cell-less architecture may decide to dynamically tune the transmission power of the base stations in order to provide a wider connectivity and offload the base station burden.

Our proposed service-tailored TDD cell-less architecture mostly focuses on the latter feature, evaluating the benefits and drawbacks of such implementation within a C-RAN context.

III. Modelling a TDD Cell-Less system

The service-tailored TDD cell-less architecture relies on a centralized decision process in charge of creating, maintaining and configuring the overlapping multi-cell area to efficiently satisfy specific service requirements. We model user behaviours within an overlapping area when multiple base stations are ready to transmit.

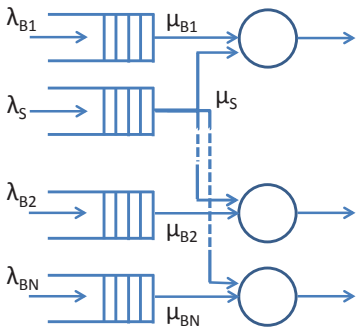


Fig. 2: Queueing system for N base stations sharing a multi-service area.

We assume that only one scheduler per base station is deployed and traffic request arrivals are exponentially distributed along the whole system coverage. Let us assume a generic scenario covered by two base stations (BS_1 and BS_2), which have an overlapping coverage area. Users are randomly distributed keeping the same position for the entire user life cycle, i.e., no mobility is assumed. Only users in the shared region, namely *multi-service area*, of Fig. 1 are simultaneously connected with both base stations, exploiting the novel frame structure, called *cell-less frame*. In principle, we rely on the idea of a multi-connectivity protocol stack, where multiple physical layer are managed by one common MAC layer, in charge of instructing the traffic on the most appropriate physical layer channels, as detailed in [10]. Multi-connectivity operations are completely transparent to the user viewpoint, leaving the burden of physical layer interface switch to the C-RAN controller. Those users, called *dual-users*, are entitled to use both UL and DL resources from different cells in a seamless way as long as resources are available. The rest of the cell is assumed to run as a legacy cell, where users are connected to a single base station sharing the same UL/DL resources with the dual-users.

Our system prevents dual-users from using shared resources when legacy users are being served. This guarantees fairness between legacy users and dual-users, as the latter intrinsically obtain a higher probability of getting served by multiple base stations, which must be compensated with a priority mechanism to avoid legacy users starvation. Therefore, we define a priority p_c per user-class c given that dual-users experience a lower priority than legacy users, i.e., $p_s < p_l$, where $l \in \mathcal{N}$ is the base station index while $N = |\mathcal{N}|$ is the total number of base stations in the system. Therefore, dual-users can seamlessly utilize resources from the attached base stations, unless a legacy user is occupying them.

We model this network behaviour and provide a useful framework to assess the dual-users QoS in terms of communication delay based on different application profiles. Analyzing user arrival rates, user channel conditions and the total amount of base stations sharing the multi-service area, our framework provides, in the next section, a good approximation of the waiting time for dual-users by means of queueing theory.

A. Queueing system model

Users join the network according to the user distribution function $\Lambda(x, y)$. We can identify different arrival rates λ_l related to specific area sections $l = 1, 2, \cdots, N$ as well as the arrival rate λ_s of dual-users in the multi-service area. Whenever an incoming user reaches the system, it sends a DL or UL request to the C-RAN baseband unit, which is in charge of deciding the resource allocation scheme for each particular traffic request. The user populates the queue based on the area it resides, and thus, it is being served according to predefined queue priorities.

A simple queuing model describing the system is depicted in Fig. 2, where single queues are specified for any cell in the system, whereas the shared queue in the middle refers to the shared multi-service area. Only N servers are listed as N base stations may accommodate traffic requests in the system. Queues B_i have a greater priority than queue S. No need to define priority levels between queues B_i as they are not served by the same server. We assume that the service rate is a i.i.d. random variable drawn from an exponential distribution with an average value equal to $\mu_l = E[\mu^{(k)}]$. The service rate directly depends on the channel quality, and in turn, on the user throughput and Signal-to-Interference-Noise-Ratio (SINR) experienced by each user joining the network and asking for traffic [11]. A different average service rate might characterized a specific area. Interestingly, the average service rate for multi-service area experiences lower values than the rate assigned to the legacy base station area, where holds the following condition $\mu_s < \mu_l$, for $l = 1, 2, \cdots, N$. The rationale behind is that multi-service area is usually placed between two or more base stations. Hence, the average distance of a random user dropped in that area is likely greater than an average distance of a random user dropped in the legacy base station area, resulting in a lower SINR, and thus, in lower service rate for the multi-service area.

As soon as a generic user joins the queue S, he will be next served if no other users are waiting in queue S and whether i) no users are waiting in queue B_1 OR ii) if no users are waiting in queue B_2. Those conditions might be summarized in the following statement.

Let us consider only two base stations. We define the residual time that an ongoing request takes to complete and exit the queue as T_0. Also, we define as T_M the time needed to serve all the other users arrived before the reference user in the same queue S. Lastly, we defined as T_Z^i the time needed to empty [1] queue B_i, where $i = \{1, 2\}$. The waiting time T_S of the reference user in the queue S will be derived as follows

$$T_S = T_0 + T_M + \frac{1}{N}\sum_{i=1}^{N} T_Z^i, \tag{1}$$

where $T_M = \frac{\lambda_S}{\mu_S} T_S$ and $T_Z^i = \frac{\lambda_{\mathrm{B}i}}{\mu_{\mathrm{B}i}}(T_i + T_S)$. We define the queue utilization as $\rho_i = \frac{\lambda_i}{\mu_i}$. Thus, we obtain

[1] This is the time to serve all users waiting in a queue with a greater priority plus the time to serve all users arriving right after the reference user, which will be served priorly.

$$\begin{aligned} T_S &= T_0 + \rho_S T_S + \frac{1}{N}\sum_{i=1}^{N} \rho_i (T_i + T_S) \\ &= T_0 + \rho_S T_S + \frac{1}{2}\left(\rho_{\mathrm{B}1}(T_{\mathrm{B}1} + T_S) + \rho_{\mathrm{B}2}(T_{\mathrm{B}2} + T_S)\right) \\ &= T_0 + \rho_S T_S + \frac{1}{2}\left(T_S(\rho_{\mathrm{B}1} + \rho_{\mathrm{B}2}) + T_{\mathrm{B}1}\,\rho_{\mathrm{B}1} + T_{\mathrm{B}2}\,\rho_{\mathrm{B}2}\right) \\ T_S &= \frac{2\,T_0 + \sum_{i=1}^{N} T_{\mathrm{B}i}\,\rho_{\mathrm{B}i}}{2 - \sum_{i=1}^{N}(\rho_{\mathrm{B}i} + \rho_S)}, \end{aligned} \tag{2}$$

where the residual time is obtained through [12] as $T_0 = \sum_{i=1}^{N} \frac{\lambda_i \hat{\mu}_i}{2}$, where $\hat{\mu}_i$ is the second moment of the statistical distribution $\mu^{(k)}$. When there is no utilization for queues B_1 and B_2, it holds the following relation

$$T_S = \frac{T_0}{1 - \rho_S}, \tag{3}$$

showing that the time of being in queue S only depends on the arrival rate λ_S and service rate μ_S of such queue, as the others queues are empty.

We can generalize the model for multiple cell surrounding the multi-service area, but we omitted detailed results for space limitation reasons. For this general case, the waiting time for a joining user in the multi-service area is defined as

$$T_S = \frac{N\,T_0 + \sum_{i=1}^{N} T_{\mathrm{B}i}\,\rho_{\mathrm{B}i}}{N - \sum_{i=1}^{N}(\rho_{\mathrm{B}i} + \rho_S)}. \tag{4}$$

Intuitively, the system is stable, e.g, no starvation, only if the following condition holds

$$N\,\rho_S < N - \sum_{i=1}^{N} \rho_{\mathrm{B}i}. \tag{5}$$

Given a certain application requirement A defined as maximum delay D_A and assuming that all legacy queues are stable, e.g., $\rho_{\mathrm{B}i} < 1$, and the shared queue experience a finite waiting time, e.g., equation (5) holds, we can calculate the probability that a generic user running application A may experience service disruption as

$$\Pr\{T_S > D_A\}. \tag{6}$$

Therefore, we can derive the system conditions by properly modelling arrival and service rates for the areas above identified and, hence, adjust the network parameters accordingly.

B. Power level optimization in the multi-service area

The average time T_S of users served in the multi-service area depends on the utilization factors ρ_S as well as on the utilization factors of the involved base stations $\rho_{\mathrm{B}i}$. In turn, the utilization factor is related to the arrival rate λ_S and the serving rate μ_S. While the former can be easily derived by the

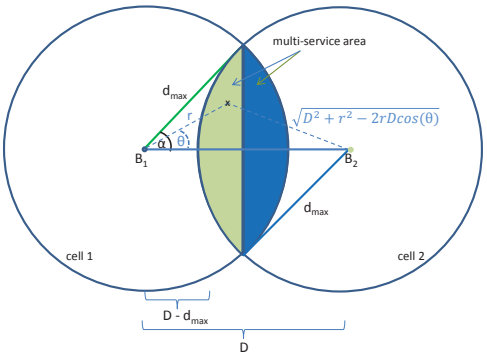


Fig. 3: Geometrical analysis of the proposed cell-less solution.

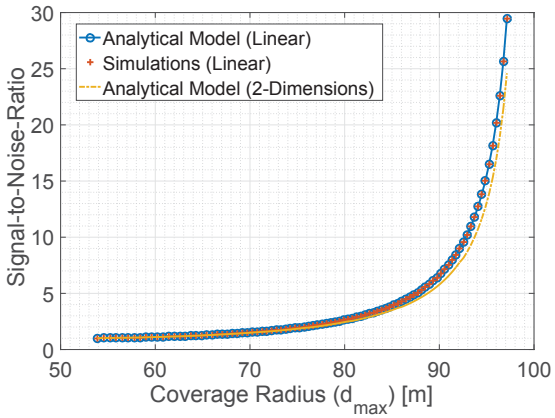


Fig. 4: Analytical against experimental results with 2 base stations placed at 100 m distance.

geometrical scope of a single area, the latter can be analytically computed as the average spatial efficiency of a potential user placed within the considered area.

Let us consider the scenario depicted in Fig. 3, where only two base stations are deployed. We aim at calculating the average bit efficiency for the multi-service area as a function of the transmitting power. We rely on the assumption that base stations transmit at the same power level, which may be dynamically adjusted to provide a larger multi-service area capability. $d_{\max}$ represents the maximum coverage radius based on the antenna sensitivity. Each user experiences a signal-to-noise-ratio equal to $\frac{P \cdot \text{PL}}{\sum_{j \in \mathcal{N}} I_j \cdot \text{PL} + N_0}$, where N_0 is the background noise while PL is the pathloss based on the propagation model, which is inversely proportional to the square of the distance [2].

Given that the base stations are transmitting with the same power levels, i.e., $S = I \cdot \text{PL} = P \cdot \text{PL}$, and the background noise is much less than the transmitting power, e.g., $N_0 \ll S$, we can rewrite the SINR for all the users residing on a spatial line connecting the base stations, as follows:

$$\text{SINR} \approx \frac{S \cdot x^{-2}}{S(D-x)^2} \tag{7}$$

where x is the distance from the serving base station and D the inter-site distance between the base stations. Thus, the average spatial SINR is calculated according to the following equation

$$\hat{\text{SINR}}_{\text{lin}} = \frac{1}{2d_{\max} - D} \int_{D-d_{\max}}^{d_{\max}} \frac{S \cdot x^{-2}}{S(D-x)^2} \mathrm{d}x. \tag{8}$$

The average spatial bit efficiency $\hat{\mu}$ is obtained by solving Eq. (8) and applying the Modulation-and-Coding-Scheme (MCS) function $f_{\text{MCS}}(\cdot)$ as follows

$$\hat{\mu} = f_{\text{MCS}}\Big(\frac{(D-d_{\max})^3 - d_{\max}}{d_{\max}(D-d_{\max})} + 2(2(D-d_{\max}) - D) - 2\,D\,\log\Big(\frac{D-d_{\max}}{D}\Big)\Big). \tag{9}$$

The more the transmitting power, the larger the multi-service area, the higher the $d_{\max}$ value. In Fig. 4, we report the SINR curve while increasing the power, e.g., the coverage area $d_{\max}$. Results are counter-intuitive: a larger transmitting power level brings a higher average spectral efficiency, hence users are served with a higher MCS. The rationale behind relies on the larger multi-service area with more users located closer to the serving base stations. Therefore, as described in the next section, a larger multi-service area results in a higher service rate μ_S as well as a higher arrival rate λ_S but exhibiting different trends, which directly impact on the performance of our queueing model.

Let us now extend our model to a 2-dimensional model, where users are covered within a maximum distance $d_{\max}$ as to create a circle surrounding the base station, as shown in Fig. 3. We want to evaluate the SINR experienced by a pilot user within the multi-service area depicted as the intersection between the two circles, i.e., all users placed within $d_{\max}$ distance from BS_1 and BS_2.

Recalling the Carnot theorem, we can write the SINR for any evaluated point in polar coordinates (r, θ), assuming the same transmitting power level is set, as follows

$$\text{SINR} = \frac{S(r^{-2})}{S(r^2 + D^2 - 2Dr\cos(\theta))^{-1}} \tag{10}$$

where r is the distance from the serving base station whereas θ is the angle between the line connecting the considered user position with the serving base station and the line connecting the base stations. Our aim is to calculate the average spatial SINR within the multi-service area A, e.g., the blue and the green sector as depicted in Fig. 3. According to the fundamental theorem of calculus, we can compute the average spatial SINR as follows

$$\hat{\text{SINR}}_{\text{2D}} = \frac{1}{A}\Big(\int_{-\alpha}^{\alpha} \int_{\frac{D/2}{\cos\theta}}^{d_{\max}} \frac{r^2 + D^2 - 2D\cos\theta}{r^2}\,\mathrm{d}r\,\mathrm{d}\theta + \tag{11}$$

$$\int_{-\frac{\alpha}{2}}^{\frac{\alpha}{2}} \int_{\frac{d_{\max}+D/2}{\cos\theta}}^{2(d_{\max}\cos\theta} \frac{r^2 + D^2 - 2D\cos\theta}{r^2}\,\mathrm{d}r\,\mathrm{d}\theta\Big). \tag{12}$$

where α is the angle obtained by drawing a segment between the crossing-point of the coverage areas and the serving base station. We can calculate α as

$$\alpha = \cos^{-1}\Big(\frac{D}{2d_{\max}}\Big). \tag{13}$$

Both extended model and linear model exhibit the same behaviour perfectly in line with our simulation results, as shown in Fig. 4. This makes our models suitable for advanced analysis as they can be easily numerically treated.

The serving rate μ_s for the multi-service area is then obtained as function of the transmitting power P, in the following equation

$$\mu_s = \hat{\mu} = f(P) = f_{MCS}(\widehat{\text{SINR}}_{\text{2D}}) \tag{14}$$

while the arrival rate λ_s is calculated according to the user distribution function $\Lambda(x, y)$ as follows

$$\lambda_s = \int_A \Lambda(x, y) \mathrm{d}A. \tag{15}$$

Serving rate $\mu_{\text{B}i}$ and arrival rate $\lambda_{\text{B}i}$ for the legacy coverage can be straightforwardly calculated.

Given a particular application target delay requirement, combining Eq. (6) with Eq. (14) and Eq. (15) we can optimally tune the transmission power levels with an exhaustive search as no closed-form solution is available.

IV. Performance evaluation

We carried out an exhaustive simulation campaign for a 2-base-stations deployment in a $3D \times 2D$ urban area network. We assume the log-distance pathloss as propagation model. However, our results can be easily extended to other complex models and environments. We use an inter-site distance $D = 100$m, wherein base stations are placed at coordinates $(D; D)$ and $(2D; D)$, correspondingly. Base stations transmission power levels are centrally adjusted within a range between 0.2 Watt and 1.2 Watts, resulting in different multi-service area size, e.g., from 12m to 96m wide.

A. Cell-less approach validation

We first compare our solution against an advanced access selection scheme [13], where users are associated beforehand with different base stations based on the current traffic load. We assume that a transmitting power $P = 24$ dBm is set on both base stations, corresponding to a 39m multi-service area. Users are uniformly distributed over the network but different traffic request distributions are compared in our evaluation: i) static distribution, in which users require the same amount of traffic periodically, ii) peak-distribution, in which traffic requests are Gaussian-distributed moving their peaks from one base station area to the other. We simulate a LTE-compliant system within a network simulator developed in MATLAB, where a video application generates one H.264 video flow (encoded at 440 kbps). We evaluate the distribution of incurred delays when the video application retrieves the content from the associated base station. When a cell-less architecture is in place, users within the multi-service area get the video content from either base station involved in a seamless manner.

In Fig. 5, we show the probability distribution function of the application delay in retrieving the video content. We mark with the vertical line the maximum accepted delay for such application $D_A = 100$ms, as defined in [14]. When a static distribution is considered, most of the users are able to get the

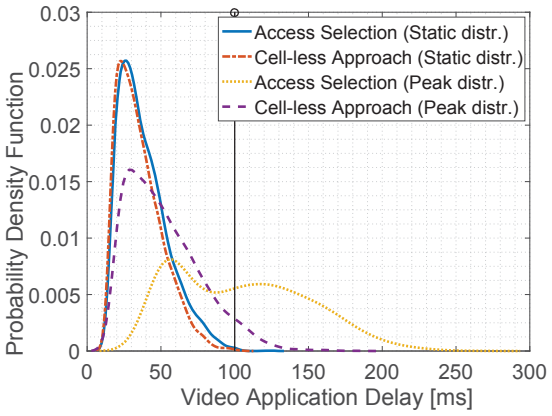


Fig. 5: PDF of video application delay.

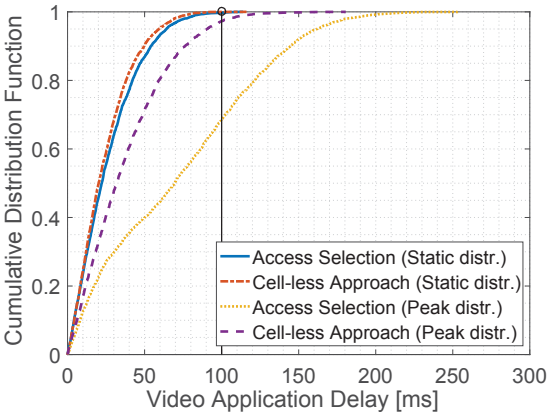


Fig. 6: CDF of video application delay.

video content within the target delay, i.e., 99.9% of the users. No relevant differences are shown between cell-less approach and access selection mechanism, due to the static nature of the traffic requests. Once we consider the peak-distribution model, user associations defined by the access-selection scheme are not able to follow the network changing. This directly affects the system performance, showing that 31% of the users do not meet the delay constraint. When we consider a cell-less approach, we see that only 2% of the users are violating the delay constraint, exhibiting outstanding results. This is also confirmed by Fig. 6, where the cumulative distribution functions are depicted. The access selection scheme exhibits an average delay equal to 72ms while the cell-less solution shows 37.19ms as average delay. Unbalanced loads are efficiently handled by the cell-less approach, as the users may take opportunistically advantage from the underloaded base station.

B. C-RAN Power adjustment

We leverage our analysis in order to design a simple mechanism to dynamically adapt the transmission power, e.g., the multi-service area coverage, by means of C-RAN capabilities. We evaluate two different traffic models, as shown in Fig. 7. While the uniform traffic model guarantees the same load along the whole network, the unbalanced traffic model mostly overloads the second base station. We evaluate the utilization factor ρ_l for any single area l, as explained in Section III-A, when different transmitting power levels are considered, i.e.,

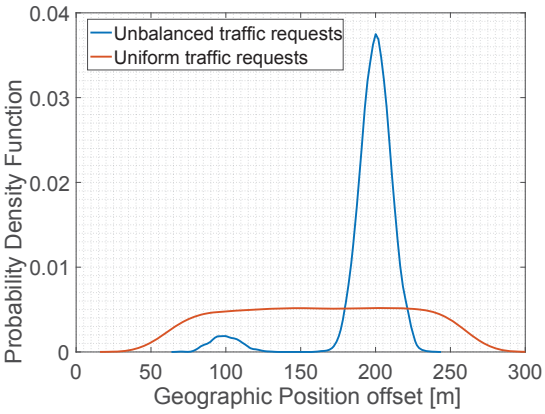


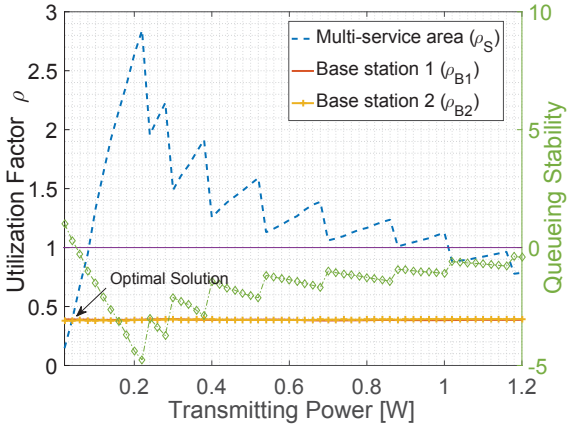
Fig. 7: Distributions of different traffic profiles.

different multi-service area sizes.

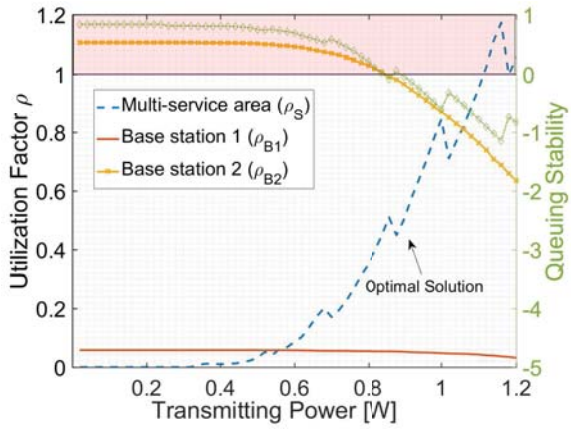
In Fig. 8, we show different utilization factors based on the considered traffic model. In addition, we show on the right y-axis, the stability equation (5) of our queueing model. A negative value corresponds to a infinite waiting time, i.e., starvation for users within the multi-service queue. As shown in Fig. 8(a), base stations are able to manage the traffic requests without requiring the multi-service area. Interestingly, increasing the transmitting power results in the same utilization factor $\rho_{\mathrm{B}i}$ as the serving rate $\lambda_{\mathrm{B}i}$ is equal to the maximum value achievable, while the arrival rate $\lambda_{\mathrm{B}i}$ does not change. This is due to the system topology, as the more transmitting power, the more cellular coverage, the more multi-service area size. Conversely, in the multi-service area the arrival rate λ_S grows with a different pace with respect to the serving rate μ_S. This leads to an increasing utilization factor ρ_S and thus, to a system instability. We highlight the optimal solution for $P = 20$ dBm, e.g., when no multi-service area is built. Nonetheless, the multi-service area is needed when the second base station is overloaded, as depicted in Fig. 8(b). While the transmitting power levels increases, the utilization factor for the second base station ρ_{B2} decreases, exhibiting the optimal solution at $P = 29$ dBm. Moving to the right of the optimal solution leads the system to instability, as the multi-service area queue is getting overloaded.

V. Conclusions

In this paper we have cast the concept of service-tailored TDD cell-less approach into the C-RAN architecture. We have shown the main benefits of providing users with multi-connectivity capabilities, which enable them to opportunistically use resources from multiple base stations. We have proposed a powerful framework to dynamically adjust the transmission power of C-RAN controlled base stations in order to satisfy specific application delay requirements. Average delay information about cell-less users are retrieved by means of queueing system concepts. Lastly, an exhaustive simulation campaign has validated our hypothesis and has proved that significant performance improvements can be achieved against a classical access selection single-connectivity solution.



(a) Balanced traffic load.



(b) Unbalanced traffic load.

Fig. 8: Utilization factor analysis for different traffic profiles.

References

[1] "GSMA, The Mobile Economy, 2016."

[2] 3GPP TS 36.213, "Physical layer procedures," (Release 13) v13.1.1, Mar 2016.

[3] H. Ishii et al., "A novel architecture for LTE-B :c-plane/u-plane split and phantom cell concept," in *GC Workshops*, Dec 2012, pp. 624–630.

[4] A. Gudipati, D. Perry, L. E. Li, and S. Katti, "SoftRAN: Software defined radio access network," in *ACM SIGCOMM Workshop HotSDN*, 2013, pp. 25–30.

[5] R. Riggio, K. Gomez, L. Goratti, R. Fedrizzi, and T. Rasheed, "V-cell: Going beyond the cell abstraction in 5g mobile networks," in *Network Operations and Management Symposium (NOMS)*, May 2014, pp. 1–5.

[6] K. Samdanis, R. Shrivastava, A. Prasad, P. Rost, and D. Grace, "Virtual cells: Enhancing the resource allocation efficiency for TD-LTE," in *Vehicular Technology Conference (VTC Fall)*, Sept 2014, pp. 1–5.

[7] K. Samdanis, R. Shrivastava, A. Prasad, D. Grace, and X. Costa-Perez, "TD-LTE virtual cells: An SDN architecture for user-centric multi-eNB elastic resource management," *Computer Communications 2016*.

[8] S. Costanzo et al., "An SDN-based virtual cell framework for enhancing the qoe in TD-LTE pico cells," in *IEEE QoMEX*, May 2015.

[9] C. Vitale, V. Sciancalepore, A. Asadi, and V. Mancuso, "Two-level opportunistic spectrum management for green 5G radio access networks," in *IEEE OnlineGreenComm 2015*, pp. 78–83.

[10] A. Ravanshid and et al., "Multi-connectivity functional architectures in 5G," in *IEEE ICC Workshop (5G-Arch)*, May 2016.

[11] X. Jin and G. Min, "Modelling and analysis of priority queueing systems with multi-class self-similar network traffic: A novel and efficient queue-decomposition approach," *IEEE Transactions on Communications*, vol. 57, no. 5, pp. 1444–1452, May 2009.

[12] L. Kleinrock, *Queueing Systems Theory 2*. Wiley-Interscience, 1976.

[13] S. Lee et al., "Vertical handoff decision algorithms for providing optimized performance in heterogeneous wireless networks," *IEEE Transactions on Vehicular Technology*, vol. 58, pp. 865–881, Feb 2009.

[14] 3GPP TS 23.303, "Policy and Charging Control Architecture," (Release 13) v13.7.0, Mar 2016.