Service-Aware User-Centric Clustering and Scheduling for Cloud-RAN with Coordinated Multi-Point Transmission

Anthony Beylerian* and Tomoaki Ohtsuki†
* Graduate School of Science and Technology, Keio University, Yokohama, Japan
E-mail: anthonybeylerian@ohtsuki.ics.keio.ac.jp
† Department of Information and Computer Science, Faculty of Science and Technology, Keio University, Yokohama, Japan
E-mail: ohtsuki@ics.keio.ac.jp

Abstract—**In this paper, we investigate the dynamics of Coordinated Multi-Point (CoMP) in a Cloud Radio Access Network (C-RAN) deployment. Cloud Radio Access Network (C-RAN) and Coordinated Multi-Point (CoMP) are two of different solutions currently being evaluated for the next generation of mobile networks (5G). In this context, our focus is on the clustering of transmitters and time-frequency resource scheduling, in systems utilizing CoMP in a C-RAN deployment. We propose a service-aware user-centric scheme for the downlink. This approach relies on adaptively creating overlapping clusters on a per-user basis and jointly scheduling the users in their preferred CoMP set. Moreover, resource scheduling is achieved with both time and frequency domain considerations, but with a fixed value in the power domain. We show that with the proposed scheme, throughput and delay improvements can be achieved for both center and edge users in a fair fashion.**

I. Introduction

With the recent proliferation of Internet services, as well as the remarkable adoption of mobile devices, mobile networks are constantly being pushed to their limits. In contrast to previous generations, it is expected that the next generation will be even more data-driven. Effectively, much effort is made by both industry and academia to provision for this increasing demand. A popular design for the new generation architecture consists of a centralized architecture known as Cloud Radio Access Network (C-RAN), as well as its enhanced version Advanced C-RAN [1], where distributed radio units are managed through centralized (regional) controllers, along with the respective baseband processing units (BBU pool). This new architecture is expected to assist and complement the network back-haul to accommodate the growing demand.

On the other hand, new concepts on all layers of the architecture have also emerged in the industry such as Self-Organized-Networks (SON) [2], HetNets [3], Phantom Cell [4], Carrier Aggregation [5], massive and distributed MIMO [6] and so on. In most cases, coordination techniques are seen as the required enablers and are being actively discussed in the literature. Coordinated Multi-Point (CoMP) is one of such coordination techniques that has evolved from Distributed Antenna Systems (DAS) and is gaining increasing interest due to its many expected advantages. In fact, recent market solutions have also been adopting C-RAN and CoMP techniques even for indoor scenarios such as in [7]. The main idea of CoMP is to orchestrate the joint transmissions and receptions from multiple source/destination stations through coordinated scheduling and/or beam-forming (CS/CB) as well as joint processing (JP). Each CoMP approach has its own merits when it comes to comparing performance gains, however, because of many practical limitations, usually straightforward and reactive schemes relying on statistical channel information are preferred for robust coordination. Moreover, the mentioned stations are usually referred to as transmission points (TPs) in the downlink, and it is envisioned to represent any class of serving station (macro, micro, pico, femto etc.) for which coordination can be applied, making CoMP an interesting and flexible solution.

II. Motivations and Related Work

Recent surveys on CoMP such as [8], mention that although different approaches have been proposed, more research is required for dynamic cell clustering, as well as opportunistic and preferably robust and low complexity scheduling in the CoMP paradigm. There are different dynamics to consider in the C-RAN architecture and clustering can be achieved differently on several layers. For our study we focus on the MAC-PHY layers and the clusters defined are strictly related to frequency resource management.

Effectively, dynamic cell clustering techniques have been previously proposed by different researchers, and usually rely on optimizing certain targets such as geometry gain or goodput [9]. However, finding the best clusters optimizing these metrics should not be treated separately from resource scheduling, since the lack of resources from a selected TP would render that TP useless in the serving CoMP set. Although many researchers treat each problem separately, others have worked on joint approaches such as [10]. In fact, authors in [10] proposed to maximize the achievable rate by grouping users in three different ways and then scheduling them on a cluster-basis following a proportional fairness (PF) rule. Although this approach shows gains over static clustering, it is not fully user-centric since it does not allow for overlapping

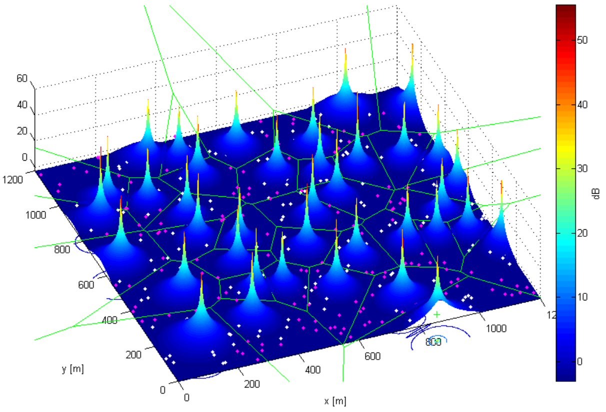

Fig. 1. Isolated cells downlink SIR.

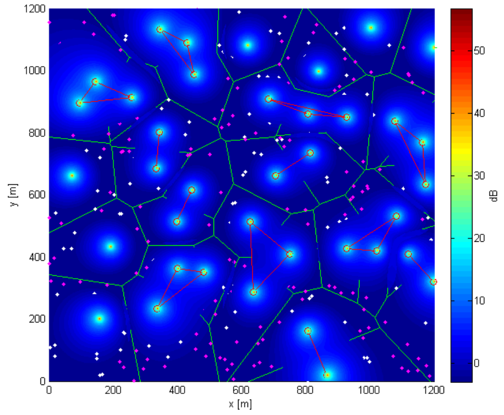

Fig. 2. Station clusters downlink SIR.

clusters and does not take into consideration the service type. Practically, user bearers are not homogeneous, meaning that users have different quality of service (QoS) requirements depending on the higher layer application (i.e. Voice, Web, Video, Gaming).

Otherwise, authors in [11] focus on the power domain, where they consider the cooperative multi-user MIMO scheme (CO-MU-MIMO). In this case, the approach is specifically designed for a hexagonal setup, where long term cooperation regions are defined based on efficiency metrics and pre-determined clustering patterns were used. This was later extended by [12] where the authors propose to achieve the CoMP-MIMO operation (up to three users) following the computation, for all resource blocks, of four metrics, of which the maximum determines the final operation. In [12], different patterns for the same cell can exist however, the patterns are split in the frequency domain, so each pattern can only exist in one sub-band, which is not always the best case. In this study, we consider single-user joint-transmission CoMP (SU-JT-CoMP), since it simplifies the analysis, however the approach can also be extended to CO-MU-MIMO to get additional multiplexing in the power domain.

In any case, if the CoMP sets are not adaptively chosen in a user-centric fashion, the scheduled resource block (RB),TP combination to serve disadvantaged users such as edge users or ones with higher pathloss, would not always yield the best results as the probability of outage could still be high. This is previously argued by authors in [9], who show that a UE-centric solution will be optimal in terms of both outage probability as well as throughput (so called good-put), although they only discuss it from a clustering perspective but propose to extend their work with a distributed graph coloring scheme for scheduling (however this might suffer from high complexity with a large number of colors).

Besides, the scheduling approach should give higher priority to bearers with more stringent requirements on higher layers. For example, bearers carrying HTTP or FTP can allow for more delay by differing their resource reservations in the time domain, in case other bearers are competing for resources. In addition, some services require a guaranteed bit rate (GBR) to achieve the required quality of experience for the users. Therefore, the scheduling should also be service-aware as to be able to consider each user bearer's QoS class.

Other recent proposals try to model the problem as a cell muting problem. For example, in [13] authors compared distributed and centralized solutions for cell muting, however they consider only static clustering for co-located servers. The problem with cell muting is that some TPs do not use all the available resources efficiently since they are powered off on the muted resources.

Additionally, authors such as in [14] proposed a PF approach using message passing, however this is not optimal in C-RAN because it does not benefit from all of the available network information as well as could become more complicated when considering inter-station communication delays.

Because of this, we propose a service-aware user-centric clustering and scheduling scheme that would enhance the user experience in the downlink. It attempts to achieve the following merits : the scheduling/clustering is fully UE-centric, relies on existing signaling mechanisms, follows fairness rules in both time and frequency domains for both center and edge users, load balances implicitly, and is service-aware as it supports multiple QoS classes, as well as GBR traffic. Performances are then evaluated in three operational scenarios, which are the isolated cell scenario, static or fixed clustering as well as the proposed dynamic user-centric clustering.

III. Study Model

A. *Remote Radio Units*

To simulate station locations, we chose to model the serving stations' locations in a square area using a Matrn hard-core type II point process (M-HCPP-II) [15]. Point processes are stochastic tools that can be used to model the random distribution of points in an multidimensional space. Briefly, a Poisson point process (PPP) is a point process following a Poisson distribution, which can be characterized by its dimensionality, bounds and density. Due to its elegant properties and its tractability, PPP has been increasingly used in wireless network analysis, often when simulating node locations in 2D space [15] and are used by many studies to capture the randomness in the actual networks. However, one shortcoming of PPP is that it does not account for a minimum Inter-Site Distance (ISD) between individual nodes, which is of practical importance, related to the technical and economical constraints during site planning and actual deployment. To this end, repulsive processes are needed to enforce this distance during simulations. These processes are also often called hard-core processes, owing to the hard-core distance between the different points. In fact, the M-HCPP is a biologically inspired child process of the PPP, which imposes this repulsion during point generation. In this type of process, the constraint on the ISD is enforced by conditional thinning, and can be achieved by three approaches as described in [15].

B. Subscribers

Subscriber user-equipment (UE) locations can be simulated using a regular PPP, due to the fact that a minimal distance between users cannot be expected. Therefore, these positions can be assumed as following a 2D-PPP. The subscribers' and server nodes' locations as well as the downlink (DL) signal to interference ratio (SIR) of the reference signals (RS) for from each server, can be visualized in the simulation space as in Figs. 1 and 2, to validate the coverage and spatial distributions. Edge users (magenta dots) and center users (white dots) are naturally all contained in the simulated coverage area (Voronoi cell) of their respective cell. Note that the Voronoi tessellation (green edges) overlaid in Fig. 1, delineating the DL coverage borders, is only valid when the same transmit power levels at each server station are used. If different transmit powers are used, the DL coverage areas will not respect the observed Voronoi tessellation. As for the coverage in Fig. 2, it does not follow the same tessellation since clustered stations (red edges) will transmit a common reference signal for multi-point operation, however the same tessellation is kept for reference. From this we can visually observe the multi-point coverage area being enhanced on the cell edges defined in the isolated scenario.

Having a static cluster, however, is sub-optimal as the desired operation would be to have the edge areas enhanced whenever a user is active in its relative region of interest and since we do not consider beam-forming, this can only be achieved in dynamic user-centric clustering, where the cluster is "centered" around each user. Therefore, in the latter case, we would have different overlapping clusters on each time-frequency pair. To note, users are classified as edge users if they are in the SIR hysteresis region, which means that the difference between their maximal experienced SIR from their anchor and the second highest experienced SIR from another station (not belonging to their cluster in clustered scenarios) is less than a hysteresis threshold.

C. Antenna and Propagation Model

For antenna configurations, for the sake of simplicity we consider 2D-omni SISO antennas, as improvements are assumed to increase with other configurations when providing extra diversity. As for the propagation model, we used large-scale fading with varying line-of-sight (LOS) and non-line-of-sight NLOS path-losses between the TPs and the users [16].

D. Traffic Model

To model more realistic conditions, we need to simulate different application traffics. Next Generation Mobile Networks (NGMN) group recommends using a traffic mix [17] where FTP, HTTP, Video streaming, VoIP and Gaming services are simulated for more accurate evaluation. We use this model for our simulations with similar proportions. As for packet drops, a packet is dropped from the UE buffer if its time in the buffer is larger than the maximum timeout value, defined in the standard QoS table [18].

IV. Assumptions and Proposed Scheme

A. UE Operation

In the traditional isolated cell scenario, each UE connects to its anchor station and reports its RS measurements following the standard signaling mechanisms. Effectively, to chose its anchor, each UE averages SIR measurements over a certain time window and then chooses the anchor based on the largest experienced value. In the static clustering scenario, serving stations are clustered once and those clusters do not change. The clustering rule can vary i.e. from considering co-located stations, using path-loss, or can be done manually by a network planning team. As for our study, we consider the popular rule based on coupling loss, which represents the experienced path-loss between stations. This means that stations that experience the largest averaged estimated path-loss between each other that is lower than a threshold, are clustered together. An example of this for a maximum cluster size of three is shown in Fig. 2 where clustered TPs are connected with red edges. In this case, when a UE finds its anchor, it joins its fixed cluster, and later reports SIR measurements for each station in that cluster.

As in the user-centric clustering scenario, different clusters are formed for each user, and this is argued in [9] to be the best clustering approach in terms of good-put. In this study, we use a similar mechanism and consider UE reports based on a relative SIR threshold rule, meaning only measurements related to stations that are part of the user's CoMP set are reported to the central control unit (CCU). The CoMP set represents the stations for which the measured SIR is larger than $\epsilon \times SIR^{1st}$, where ϵ is a scaling factor and SIR^{1st} is the largest experienced SIR. This generates UE-specific clusters and allows fast cell-selection, making the handover process similar to a soft-handover but based on station updates in the CoMP set. Moreover, the user reports are sent through a control channel on the strongest link. In our study, these dynamic measurement reports are stunted to the three highest measurements to compare with the fixed clustering case, and assume a reasonable limit on the control traffic overhead.

B. Scheduling Operation

In the standard scenario, the scheduling is achieved per station. As for the coordinated scenarios, the scheduling is centralized at the CCU, which collects the UE measurement reports. The only difference between the coordinated scenarios is in the clustering behavior (i.e. static or dynamic). The proposed scheduler follows the time domain-frequency domain (TD-FD) approach to achieve fairness in both dimensions. To mention, the system energy efficiency should also be analyzed since we are using multiple TPs, however for the purpose of this study we only develop on the scheduling part. Nevertheless, link adaptation is assumed in terms of modulation rate.

The general process for this is described in the simplified flow chart shown in Fig. 3. The scheduling process starts by updating the TD metric (τ) for each active UE bearer (u) at time (t) based on the service type specified by its class weight (QoS), which is the inverse of the service priority. The TD metric also depends on the maximum allowed delay (Δ) per service, the average historical rate ($\overline{R}$) and average buffer wait time ($\overline{\delta}$) :

$$\tau(u,t) = \frac{QoS(u)}{\overline{R}(u,t)} \exp\left[\beta \frac{\overline{\delta}(u,t)}{\Delta(u)}\right], \tag{1}$$

where

$$\begin{aligned}\overline{R}(u,t) &= (1-\alpha)\overline{R}(u,t-1) \\ &+ \begin{cases}\alpha\hat{R}(u,t-1), & \text{if u} \notin U(t-1) \\ 0, & \text{otherwise.}\end{cases}\end{aligned} \tag{2}$$

With $\alpha \in [0,1]$ a smoothing factor, $\beta \in [0,1]$ a weight to determine how strongly the average delay exponentially affects the metric, U is the set of scheduled user bearers, and $\hat{R}$ is the user's estimated instantaneous rate defined by :

$$\hat{R}(u,t) = K(u,t) \times W \times \sum_{k=1}^{K(u,t)} \log_2(1+\hat{\gamma}^k(u,t)), \tag{3}$$

where K is the number of allocated resource blocks (RB), W the RB bandwidth and $\hat{\gamma}$ is the SIR estimated per RB (k). The TD metric represents the user's priority in the scheduling process, therefore the list remains sorted at each update. Users with GBR bearers need to achieve at least their target bit rate, but also should not be allocated more resources than required since that would over-allocate resources that would better serve other bearers. To provision for this, we can add an exponential weight to the metric based on the average rate ($\overline{R}$) and target rate (R_{GBR}). Also, if μ is a binary variable representing the condition that u has a GBR bearer, the final metric becomes:

$$\tau(u,t) = \frac{QoS(u)}{\overline{R}(u,t)} \exp\left[\beta \frac{\overline{\delta}(u,t)}{\Delta(u)} + \mu(u)\rho\Big[1 - \frac{\overline{R}(u,t)}{R_{GBR}(u)}\Big]\right]. \tag{4}$$

where ρ is a weight similar to β but for the rate fluctuations.

Since there are different variables involved and several stages in the scheduler, finding an analytical proof for the best parameters to set is non-trivial. In fact, the choice of parameter combinations of β and ρ was done experimentally, and is discussed in the following section. However, if we look at the expression of the metrics, we can have a better idea about the dynamics involved. In the TD metric, the first ratio related to delay is an increasing function from 0 to 1. This is because packets cannot have a delay larger than the maximum allowed delay (they will be dropped). As for the second term related to the GBR rate, it is only included when the bearer is for a GBR service. However, in this case it is first a decreasing function from 1 to 0, when the average rate is lower than or equal to the GBR rate. For larger values, the term becomes negative since the rate ratio becomes larger than 1. This was designed like so in order to decrease the priority of GBR traffic that has already satisfied its target rate. In general, the priority of a GBR bearer is increased more than that of a non-GBR bearer (due to the extra positive term in the exponential) since its target rate needs to be guaranteed. However, when it does achieve it, its priority over non-GBR traffic will decrease in order to give non-GBR traffic the priority to chose its favored resource set.

Afterwards, we schedule the bearer with highest priority by first updating its FD metric per TP (r) in its CoMP set, per available resource block (k). The FD metric represents the RB preference per TP and is calculated following :

$$\phi_k(u,r,t) = \begin{cases}1, & \text{if u} \notin U_k(r,t-1), \\ \frac{\hat{R}_k(u,r,t)}{\overline{R}(u,r,t)}, & \text{otherwise.}\end{cases} \tag{5}$$

where U_k is the set of users previously served on RB k. For a cluster $C = \{TP_1, TP_2, ..., TP_M\}$, of maximum size M, we define a sub-cluster S as any combination of TPs existing in C. For all sub-clusters of user u we calculate the product of metrics (POM) on each resource block k as:

$$\Phi_k(u,S,t) = \prod_{r \in S/S \subset C} \phi_k(u,r,t). \tag{6}$$

For $M = 3$ we will have in total seven combinations to compute and then choose the one that maximizes the POM: $\arg\max_{(k,S)} \Phi_k(u,S,t)$. The POM was designed to reduce the overhead as well as to achieve implicit load balancing. In fact, overhead is an issue in CoMP based schemes and is difficult to consider during scheduling because it depends on each user's activity. However, in order to reduce the impact of high overhead, we can limit the cluster size to three TPs. Moreover, the POM will allow to choose sub-clusters from the larger CoMP set and will avoid any wasteful allocations for users that do not really need them. For example, a UE in better conditions (experienced rate) vis-a-vis two out of three of its TPs in the CoMP set on a certain frequency resource, would prefer reserving the FD slot from only the best two TPs, allowing the unreserved resources to be used by other users. Otherwise, if it has no experience with a specific TP, it

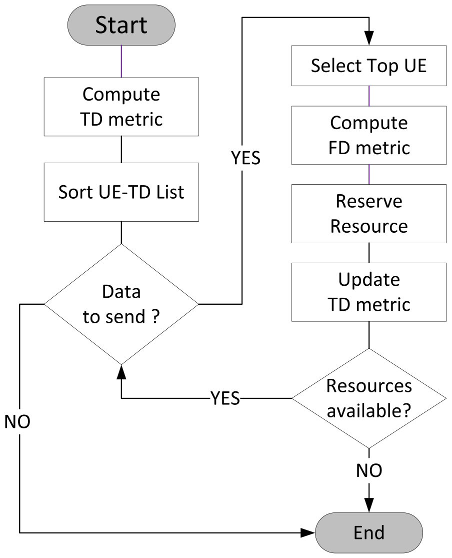

Fig. 3. Simplified scheduling process (every scheduling interval).

will give it a neutral score of one in the product and choose a sub-cluster accordingly.

Therefore, by limiting the cluster size and adding our pruning approach using the POM, we also reduce both the amount of measurement information to feedback by the users (by the upper limit) as well as the control messages to push to each station per scheduling interval. Subsequently, we update the TD metric for the chosen user with highest priority and repeat the process until either when resources are depleted or there is no more data to transmit. We use a linear update based on the number of allocated resources:

$$\tau(u,t) = \frac{\tau(u,t)}{K(u,t)+1}. \tag{7}$$

V. Performance Evaluation

We simulate the network operation with parameters summarized in Table 1, and observe the results shown in Figs 4-6. In these figures, "PF FD-TD" represents the traditional isolated cell approach using a PF scheduler augmented with our modifications for FD and TD metrics. In the "Fixed Clusters" scenario, clusters are fixed and formed based on their coupling loss. In "User-Centric Clusters", clusters are dynamically chosen in a user-centric fashion. As we can see from Figs. 4-6, the average user throughput and packet delay improve with clustering compared to a standard isolated cell scenario. Furthermore, for the dynamic approach the throughput is high compared to the fixed approach, for both center and edge users. Moreover, we also notice that with the dynamic approach, the difference in average throughputs between edge and center users is much smaller compared to the other approaches, keeping a fairer balance between both types of users, while at the same time allowing for the throughput differentiation to be only per traffic type as we can see in Fig. 4 for GBR vs. Non-GBR. This is mainly because in static clusters, we still have a cluster edge whereas in dynamic clustering, the effect of being on the cell edge is compensated by the dynamic coordination, since clusters are chosen per user.

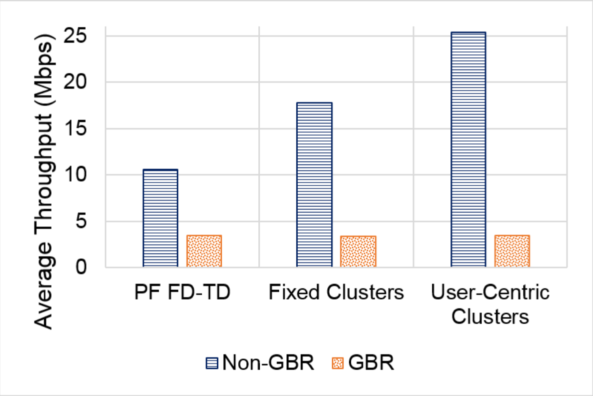

Fig. 4. Average Throughput ($\beta = 0.5, \rho = 0.5$).

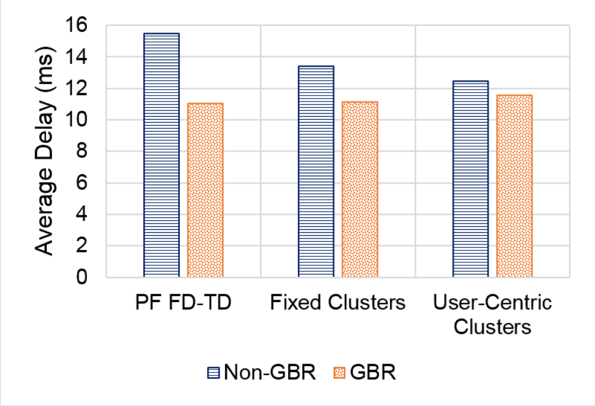

Fig. 5. Average Delay ($\beta = 0.5, \rho = 0.5$).

Furthermore, the main gain in throughput is in the non-GBR traffic as shown in Fig. 4. This is because the GBR traffic attempts to satisfy its rate requirement but then the bearer priority is decreased the more it goes higher than the rate requirement. How stringent we want this behavior to be can be set by the parameter ρ. In Fig. 5, we can see that the average packet delay is a slightly higher than 10 ms (1 frame time) and is somewhat improved in the clustering scenarios. The effect on delay can also be controlled by the β parameter. In fact, if we look at Fig. 7 where we show five representative cases, we can see that on one hand, when we give more weight to the delay fluctuations, the average delay decreases but so does the throughput. Conversely, when we increase the weight for the rate fluctuations, the throughput is improved compared to the case where the weights are the same but inversed, however it is not the best case for throughput. We can say that in this case, the delay is enhanced by sacrificing the throughput. Therefore, if a slightly higher average delay is acceptable, setting equal weights will enhance the throughput. Effectively, we have tried different values for the coefficients and have observed the best throughput gain for equal coefficients as seen in Fig. 7.

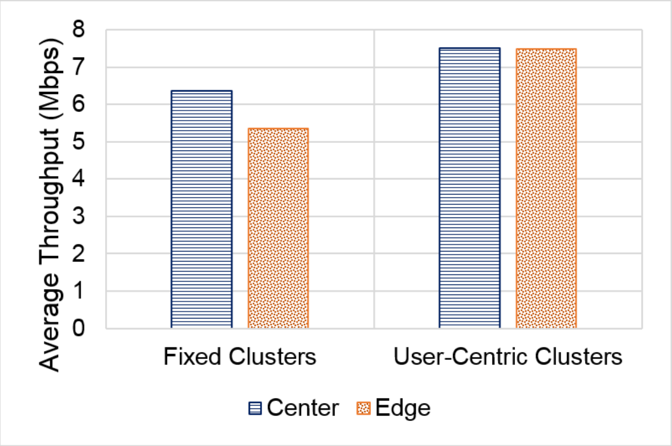

Fig. 6. Edge vs. Center Throughputs ($\beta = 0.5, \rho = 0.5$).

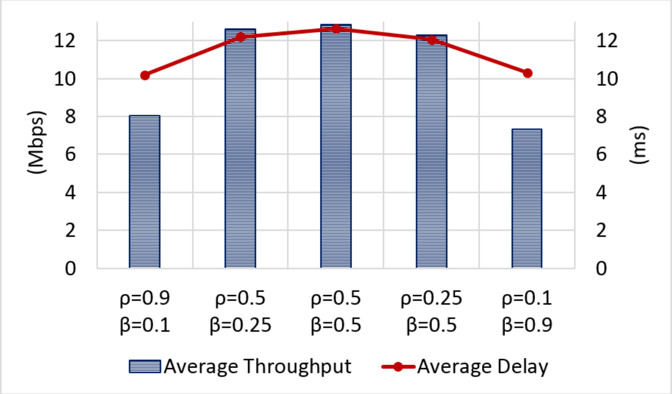

Fig. 7. Metric parameter selection.

VI. Conclusion

In this paper we have studied clustering and scheduling in the C-RAN CoMP paradigm, and have proposed a service-aware user-centric dynamic scheme for the downlink. User-centric schemes achieve the best results in terms of coverage and achievable rate. Service-awareness in scheduling must also be achieved considering that each user's activity is different in terms of the traffic type. Effectively, the approach considers both time and frequency domain perspectives while having a fixed value in the power domain, under a traffic mix of different services per user.

From our simulations, we have observed that we can expect that the proposed scheme could yield throughput improvements particularly for non-GBR traffic, while keeping the fairness between center and edge users and experiencing acceptable packet delays. However, for other practical aspects, we would still have to study and evaluate the robustness to feedback delays and sensitivity to inaccuracies in channel estimation, as well as the system energy consumption trade-off (with power domain considerations), all of which are issues that would be interesting to investigate in future work.

TABLE I
Simulation Configuration

Parameter	Simulation Model
Frequency/Bandwidth/Duplexing	2GHz/10 MHz/FDD
Region	1200 m^2
Station Locations	MHCPP-II 30.10^{-6} stations/m^2
Station ISD	80 m
Antenna configuration	2D-Omni SISO
Station power	30 W
UE Locations	PPP 300.10^{-6} users/m^2
Hysteresis Threshold	3 dB
Access Scheme	OFDMA
FFT size	1024
Scheduling Interval	1 sub-frame
Pathloss Model	3GPP Outdoor LOS-NLOS [16]
Link Adaptation	10% BLER target
Modulation Order	QPSK, 16QAM, 64QAM
Channel Estimation	Ideal
Packet Drop Time	LTE QoS table [18]
Traffic Model	NGMN mix [17]
Scheduling Weights	$\alpha = 0.7, \beta = 0.5, \rho = 0.5$
GBR threshold rate	512 Kbps
Coupling Loss Threshold	-125 dBW

References

[1] NTT DOCOMO, "DOCOMO to Develop Next-generation Base Stations Utilizing Advanced C-RAN Architecture for LTE-Advanced" *DOCOMO Press Releases*, February 21, 2013.

[2] M. Arslan et. al, "Software-defined networking in cellular radio access networks: potential and challenges" *IEEE Com. Mag.*, vol 53, no. 1,pp.150-156, 2015.

[3] O. Stanze and A. Weber, "Heterogeneous Networks With LTE-Advanced Technologies" *Bell Labs Technical Journal*, vol. 18, no. 1, pp.4158, 2013

[4] H. Ishii et. al, "A novel architecture for LTE-B : C-plane/U-plane split and Phantom Cell concept" *Globecom Workshops*, 2012.

[5] Guangxiang Yuan, Xiang Zhang, Wenbo Wang, Yang Yang, "Carrier aggregation for LTE-advanced mobile communication systems" *IEEE Com. Mag.*, vol.48, no. 2, pp. 88-93, 2010.

[6] T.L. Narasimhan et. al, "Large-Scale Multiuser SM-MIMO Versus Massive MIMO" *Information Theory and Applications Workshop*, 2014.

[7] Airvana OneCell, *http://www.airvana.com/products/enterprise/onecell/*.

[8] G.Y. Li et. al, "Multi-Cell Coordinated Scheduling and MIMO in LTE" *IEEE Com. Surveys and Tutorials*, vol. 16, no.2, pp. 761-775, 2014.

[9] V. Garcia et al, "Coordinated Multipoint Transmission in Dense Cellular Networks With User-Centric Adaptive Clustering" *IEEE Transactions on Wireless Com.*, vol. 13, no. 8, pp. 4297-4308, 2014.

[10] P. Baracca et al, "A Dynamic Joint Clustering Scheduling Algorithm for Downlink CoMP Systems with Limited CSI" *ISWCS*, pp. 830 834, 2012.

[11] N. Kusashima et al., "Dynamic Fractional Base Station Cooperation Using Shared Distributed Remote Radio Units for Advanced Cellular Networks" *IEICE TRANSACTIONS on Communications*, vol.E94-B No.12 pp.3259-3271.

[12] D. Matsuo et al., "Shared Remote Radio Head Architecture to Realize Semi-Dynamic Clustering in CoMP Cellular Networks" *IEEE GLOBECOM 2012 Workshop on Multicell Cooperation*, Dec. 2012

[13] X. Wang et al, "Coordinated Scheduling and Network Architecture for LTE Macro and Small Cell deployments" *IEEE ICC Workshop*, 2014.

[14] K. Kwak et. al, "Adaptive and Distributed CoMP Scheduling in LTE-Advanced Systems" *IEEE VTC*, 2013.

[15] H. El Sawy et. al, "Stochastic Geometry for Modeling, Analysis, and Design of Multi-Tier and Cognitive Cellular Wireless Networks: A Survey" *IEEE Commun. Surveys and Tutorials*, vol. 15, no. 3, pp. 996 1019, 2013.

[16] 3GPP TR 36.828, "Further enhancements to LTE Time Division Duplex (TDD) for Downlink-Uplink (DL-UL) interference management and traffic adaptation" *V11.0.0.*.

[17] NGMN, "NGMN Radio Access Performance Evaluation Methodology" *WhitePaper*, 2008.

[18] 3GPP TS 23.203, "Policy and charging control architecture" *Technical Specification Group Services and System Aspects*, V13.2.0.