This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JSAC.2016.2624119, IEEE Journal on Selected Areas in Communications

Exploiting Hybrid Clustering and Computation Provisioning for Green C-RAN

Kun Guo, Min Sheng, *Senior Member, IEEE,* Jianhua Tang, *Member, IEEE,*
Tony Q.S. Quek, *Senior Member, IEEE,* and Zhiliang Qiu, *Member, IEEE*

***Abstract*—By migrating baseband processing functionalities into a centralized cloud-based baseband unit (BBU) pool, cloud radio access network (C-RAN) facilitates cooperative transmission among remote radio heads (RRHs) and enables flexible computation provisioning in the BBU pool. In C-RAN, due to the high amount of data transfer from the BBU pool to RRHs through fronthauls, limited fronthaul capacity becomes a key factor when designing cooperative transmission schemes among RRHs. Meanwhile, as computational resources are provisioned to mobile users (MUs) for baseband processing in form of virtual machines (VMs) in the BBU pool, an effective VM assignment strategy is also with great significance. In this paper, we propose a framework to design a green C-RAN under the constraint of limited fronthaul capacity, where we jointly optimize hybrid clustering and computation provisioning to appropriately provide a cluster of RRHs and a VM to each MU for cooperative transmission and baseband processing, aiming at minimizing the system power consumption. The system power minimization problem is formulated as an integer non-linear programming problem, which is hard to tackle. For tractability purpose, we transform this problem to an equivalent hybrid clustering problem embedded with a series of VM assignment problems. On this basis, we firstly achieve the optimal solution for system power minimization with high computational complexity and then, a greedy algorithm is proposed to solve the hybrid clustering problem for practical implementation. Finally, simulation results demonstrate that the proposed joint optimization of hybrid clustering and computation provisioning can significantly reduce the system power consumption.**

***Index Terms*—Cloud radio access network (C-RAN), computation provisioning, cooperative transmission, hybrid clustering, limited fronthaul capacity, power saving, virtual machine (VM) assignment**

I. Introduction

Driven by the tremendous increase in mobile traffic from ubiquitous Internet access and diverse multimedia applications, mobile network architecture is trending towards centralization to facilitate centralized processing and cooperative transmission/reception. Cloud radio access network (C-RAN) is regarded as a promising architecture in fifth generation mobile network (5G) [1]–[3]. In C-RAN, computational resources are aggregated in the centralized cloud-based baseband unit (BBU) pool, and are provisioned to mobile users (MUs) for baseband processing in form of virtual machines (VMs) by virtualization techniques. Meanwhile, low-cost remote radio heads (RRHs) are densely deployed in C-RAN to be responsible for conversions between baseband signals and radio signals and also account for radio signal transmission/reception. The high amount of data transfer between RRHs and the BBU pool relies on the high bandwidth and low latency fronthauls.

The key feature of C-RAN is that RRHs and BBUs are physically decoupled, resulting a centralized cloud-based BBU pool. The centralized management of BBU pool renders information global, and enables efficient cooperation transmission/reception among RRHs. As a result, significant performance improvements through coordinated multiple point transmission (CoMP), such as coordinated beamforming (CoMP-CB) or joint processing (CoMP-JP), can be achieved in C-RAN [4]. On the other hand, virtualization in the cloud-based BBU pool can make fully use of aggregated computational resources and improve hardware utilization. For example, [5] proposed a simulation model and validated that cloud computing and virtualization techniques in C-RAN can achieve remarkable multiplexing gains in the BBU pool and improve the utilization of computational resources.

However, several new challenges arise in C-RAN. Firstly, as BBU pool provides abundant computational resources, the majority of power consumption in C-RAN is directly attributed to baseband processing power consumption in the BBU pool [6], [7]. Therefore, the emerging problem is how to manage and assign VMs to MUs to save baseband processing power consumption significantly [8], [9]. In addition, the close proximity among RRHs results in increased interference and thus, how to effectively suppress interference and reduce power consumption at RRHs plays a vital role in C-RAN [10], [11]. Last but not least, the data with high bit rates (in the order of Gbps) exchanged between RRHs and the BBU pool makes the fronthaul capacity requirement become stringent [12], [13]. Thus, there is an urgent need to address the problems regarding limited fronthaul capacity as well.

To solve the aforementioned challenges, this work jointly takes hybrid clustering and computation provisioning into consideration to appropriately select a cluster of RRHs and a VM to each MU for cooperative transmission and baseband

This work was supported in part by the National Natural Science Foundation of China (61231008, 61301176, and 91338114), 863 project (No.2014AA01A701), and 111 Project (B08038), in part by the SUTD-ZJU Research Collaboration under Grant SUTD-ZJU/RES/01/2014, the MOE ARF Tier 2 under Grant MOE2014-T2-2-002, and the MOE ARF Tier 2 under Grant MOE2015-T2-2-104.

K. Guo, M. Sheng, and Z. Qiu are with the State Key Laboratory of ISN, Xidian University, Xi'an 710071, China (e-mail: guokun@stu.xidian.edu.cn; {msheng, zlqiu}@mail.xidian.edu.cn).

J. Tang is with the Department of Electrical and Computer Engineering, Seoul National University, Seoul, Korea. He was with the Singapore University of Technology and Design, Singapore (e-mail: jtang4@e.ntu.edu.sg). The corresponding author is J. Tang.

T. Q.S. Quek are with the Singapore University of Technology and Design, Singapore (e-mail: tonyquek@sutd.edu.sg).

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JSAC.2016.2624119, IEEE Journal on Selected Areas in Communications

processing, aiming at addressing the system power minimization problem under the constraint of limited fronthaul capacity. Particularly, the system power consumption consists of power consumption in the BBU pool and that at RRHs. In the BBU pool, we consider heterogeneous VMs, which have different computation capacities and different power consumption. To save power consumption in the BBU pool, this work exploits computation provisioning to assign an appropriate VM to each MU according to its traffic arrival rate, channel condition, and quality of service (QoS) requirement. Moreover, by taking full advantage of spatial degree of freedom (DoF) and fronthaul capacity, hybrid clustering is designed to reduce power consumption at RRHs subject to the limited fronthaul capacity. To be specific, a cluster of RRHs is formed to each MU for cooperative transmission, where several RRHs (referred to as transmitted RRHs) send desired signals to the MU while some RRHs (referred to as coordinated RRHs) avoid interfering with the MU.

A. Contributions

The main contributions of this work are summarized as follows:

- First, we develop a holistic framework to encompass cooperative transmission and computation provisioning together so as to lay the foundation of high-efficiency system power saving in C-RAN. Furthermore, we minimize system power consumption while accounting for the limited fronthaul capacity. Specifically, we formulate a system power minimization problem constrained by the limited fronthaul capacity that allows for a flexible tradeoff between power consumption in the BBU pool and that at RRHs so as to make a more judicious decision on hybrid clustering and VM assignment. In addition, power consumption at RRHs includes transmit power consumption and static power consumption, which is general enough to cover the power consumption on fronthauls.
- Second, we exploit the special structure of system power minimization problem and make the problem tractable by equivalently transforming it to a hybrid clustering problem, in which a series of VM assignment problems are embedded. Subsequently, exhaustive search method is adopted to solve the equivalent hybrid clustering problem. Therein, we transform the embedded VM assignment problem to a minimum weight perfect matching (MWPM) problem and optimally solve it using the Hungarian method. However, the exhaustive search method has prohibitive computational complexity with increasing number of RRHs and MUs. Therefore, we further propose a low complexity greedy algorithm to solve the hybrid clustering problem for practical implementation.
- Third, we obtain many interesting results through extensive simulations: 1) our proposed joint optimization schemes can strike a balance between power consumption in the BBU pool and that at RRHs so that a more judicious configuration between these two parts of power consumption can be achieved for significant system power castdown. 2) the hybrid clustering can take full advantage of the limited fronthaul capacity and spatial degree of freedom so as to promote system power saving in C-RAN.

B. Related Work

C-RAN, acting as a promising 5G architecture, has received much attention recently. On one hand, many studies addressed power minimization problem in C-RAN by exploiting full CoMP-JP [7], [10], [11], where each MU is served by all active RRHs regardless of the limited fronthaul. For instance, a cross-layer optimization of clustering, beamforming, and VMs' computation capacities was proposed by [7] to minimize the overall power consumption on all components in C-RAN. By optimizing sparse beamforming to determine the clusters and transmit power, [10] minimized the network power consumption comprised of transmit power at RRHs and transport power on fronthauls. Likewise, [11] jointly considered uplink and downlink power consumption to optimize clustering and beamforming under the full CoMP-JP mode.

On the other hand, studies on power saving, considering limited fronthaul capacity also have emerged. For example, by introducing fronthaul consumption as a penalty term in the objective function, [14] compared the network power consumption of two mainstream transmission strategies, which includes compression strategy and data sharing strategy. In particular, which transmission strategy is adopted to account for the finite fronthaul depends on the placement of beamforming operation. When beamforming operation is performed in the BBU pool, compression strategy is exploited to compress the beamforming signals forwarded to RRHs in order to reduce fronthaul consumption [15]–[17]. Alternatively, with beamforming operated at RRHs, the BBU pool simply shares each MU's desired baseband signal with its serving RRHs. In this case, data sharing strategy is useful to alleviate the burden on fronthauls, with which each MU is associated with only a subset of RRHs to reduce the amount of shared data.

Data sharing strategy, called as partial CoMP-JP as well, has been widely studied in the existing researches [12], [13], [18]–[22] and has been divided into two categories: user-centric CoMP-JP and disjoint CoMP-JP. In [12], [13], [18]–[20], authors pursued various performance metrics by optimizing user-centric CoMP-JP, where each MU is served by only a selected subset of RRHs (potentially overlapping). While [21] and [22] adopted disjoint CoMP-JP and divided the entire network into non-overlapping clusters to serve MUs separately. Although partial CoMP-JP can reduce fronthaul consumption, system performance is severely degraded by inter-cluster interference [19], [21]. Based on this, [23] proposed hybrid clustering to fully exploit spatial DoF to mitigate interference so as to improve system throughput.

Unfortunately, [10]–[23] improved the system performance regardless of computation provisioning and power consumption in the BBU pool. However, [7], [24], [25] pointed out that leveraging the benefits of pooling computational resources has a significant impact on the system performance improvement. As an early work to consider power consumption in the BBU

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JSAC.2016.2624119, IEEE Journal on Selected Areas in Communications

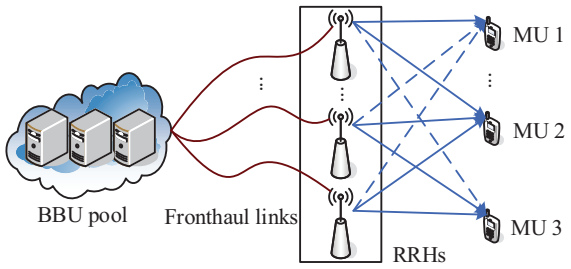

Fig. 1. An illustration of C-RAN. RRHs are termed as RRH 1, RRH 2, and RRH 3 successively from the top to the bottom. The solid and dotted lines between RRHs and MUs represent transmission links and cooperation links, respectively.

pool, [7] preassigned VMs to MUs and adjusted each VM's computation capacity to minimize system power consumption regardless of the fronthaul capacity. In fact, pricing-based VM assignment and management policies have been widely studied in cloud computing community to optimize various kinds of objectives. For instance, with multiple types of VMs taken into account, a VM assignment policy was proposed to maximize the sum of user's bid in [26]. However, these policies in cloud computing cannot be directly applied to C-RAN due to the absence study of wireless access networks.

To this end, accounting for the limited fronthaul capacity, this work aims to minimize the system power consumption comprised of power consumption in the BBU pool and that at RRHs in downlink C-RAN by jointly optimizing hybrid clustering and computation provisioning.

C. Organization

The remainder of this paper is organized as follows. In Section II, we formally describe our system model, and then the formulation and transformation of system power minimization problem are presented in Section III. In Section IV, we propose an optimal exhaustive search method to solve the equivalent hybrid clustering problem for system power minimization. Then, we further propose a low complexity greedy algorithm for the hybrid clustering problem in Section V. In Section VI, simulation results are presented and discussed. Finally, we conclude our paper in Section VII.

Notations: Boldface letters refer to vectors (lower case) or matrices (upper case). $(\cdot)^T$ and $(\cdot)^H$ respectively denote the transpose and conjugate transpose of a vector or a matrix. $\mathbb{C}^{x\times y}$ denotes the space of $x \times y$ complex matrices. $||\mathbf{x}||$ denotes the Euclidean norm of vector $\mathbf{x}$. $|\mathcal{A}|$ denotes the cardinality of set $\mathcal{A}$ and $x \in \mathcal{A}\backslash\mathcal{B}$ means that element x belongs to set $\mathcal{A}$ but not belongs to set $\mathcal{B}$. The distribution of a circularly symmetric complex Gaussian random variable with mean μ and variance σ^2 is denoted by $\mathcal{CN}(\mu, \sigma^2)$.

II. System Model

We consider a downlink C-RAN, as shown in Fig. 1. Let $\mathcal{L} = \{1, .., L\}$ denote the set of RRHs. RRH l is equipped with N_l antennas and is connected to the BBU pool via a fronthaul with limited capacity S_l. In the BBU pool, the set of VMs is given by $\mathcal{M} = \{1, .., M\}$. VM m is described by a tuple comprised of two elements, i.e., $\{\mu_m, \phi_m\}$, where μ_m and ϕ_m represent VM m's computation capacity and power consumption, respectively. Moreover, we express the set of single-antenna MUs as $\mathcal{K} = \{1, ..., K\}$.

A. Computation Provisioning

Computation provisioning aims to make a decision on VM assignment for MUs. In this work, we consider a simple VM assignment model, where each MU is served by one VM and one VM provides service for at most one MU. This model not only presents the parallel processing ability of the BBU pool but also exploits the aggregated computational resources to yield the computation diversity gain. On this basis, $x_{mk} \in \{0, 1\}$ is introduced to depict the VM assignment. Specifically, $x_{mk} = 1$ indicates that VM m is assigned to MU k. Otherwise, VM m does not provide service for MU k. Thus, the following constraints for VM assignment should be held:

$$\sum_{m\in\mathcal{M}} x_{mk} = 1, \forall k \in \mathcal{K}, \text{ and } \sum_{k\in\mathcal{K}} x_{mk} \leq 1, \forall m \in \mathcal{M}. \quad (1)$$

B. Hybrid Clustering

To address issues on limited fronthaul capacity, we utilize hybrid clustering to form a cluster of RRHs to each MU for cooperative transmission. The formed cluster of RRHs includes transmitted and coordinated RRHs, which can take full advantage of spatial DoF and fronthaul capacity to improve transmission performance. To describe hybrid clustering, a transmission matrix $\mathcal{T} = [t_{lk}]_{L\times K}$ and a coordination matrix $\mathcal{A} = [a_{lk}]_{L\times K}$ are introduced. Each element of these two matrices is either 0 or 1. If RRH l transmits desired data to MU k, $t_{lk} = 1$, otherwise, $t_{lk} = 0$. When RRH l does not transmit desired data to MU k but avoids interfering with MU k, $a_{lk} = 1$. Naturally, $a_{lk} = 0$ if RRH l also does not avoid imposing interference on MU k. Importantly, matrices $\mathcal{T}$ and $\mathcal{A}$ jointly identify the cluster for each MU. We first define the sets of transmitted RRHs and coordinated RRHs for MU k as $\mathcal{T}_k = \{l \mid t_{lk} = 1, l \in \mathcal{L}\}$ and $\mathcal{A}_k = \{l \mid a_{lk} = 1, l \in \mathcal{L}\}$, respectively. Then, the cluster for MU k, i.e., $\mathcal{C}_k$, is denoted by $\mathcal{C}_k = \mathcal{T}_k \bigcup \mathcal{A}_k$.

It is observed from Fig. 1 that clustering schemes under CoMP-CB and CoMP-JP mode, adopted by MU 1 and MU 2 respectively, are two special cases of hybrid clustering. To be specific, a cluster including only one transmitted RRH is formed for MU k under CoMP-CB mode, i.e., $|\mathcal{T}_k| = 1$, while the set of coordinated RRHs is empty for MU k under CoMP-JP mode, i.e., $|\mathcal{A}_k| = 0$. Moreover, hybrid clustering for MU k degenerates into clustering scheme under Non-CoMP mode when there is only one transmitted RRH and no coordinated RRH for MU k, i.e., $|\mathcal{T}_k| = 1$ and $|\mathcal{A}_k| = 0$. In terms of the definition on cluster, the decision on hybrid clustering is closely related to the fronthaul capacity and the number of antennas equipped by each RRH. Subsequently, we capture hybrid clustering with following details.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JSAC.2016.2624119, IEEE Journal on Selected Areas in Communications

1) Fronthaul Capacity Constraint: To account for the limited fronthaul, we define the fronthaul capacity at RRH l, i.e., S_l, as the maximum number of baseband signals forwarded to MUs on this fronthaul [8]. By using hybrid clustering, transmission links are established between (transmitted) RRH and its served MUs to forward desired baseband signals, and in the meanwhile, cooperation links are set up between (coordinated) RRH and its coordinated MUs for interference avoidance without baseband signal transfer. Specifically, the baseband signals transmitted on transmission links should be delivered to the RRH via its connected fronthaul. Therefore, the fronthaul consumption at RRH l is only involved with the number of its transmission links, and unrelated to the number of its cooperation links. Furthermore, we have the following fronthaul capacity constraint:

$$\sum_{k\in\mathcal{K}} t_{lk} \leq S_l, \forall l \in \mathcal{L}, \tag{2}$$

where the left hand standing for the total number of baseband signals forwarded to MUs on the fronthaul at RRH l, cannot be more than the fronthaul capacity S_l.

2) Zero forcing (ZF) Precoding : Based on the cluster for each MU, the received signal at MU k can be expressed as

$$\begin{aligned} y_k \quad = \quad & \sum_{l\in\mathcal{T}_k} \mathbf{h}_{lk}^H \mathbf{w}_{lk} \sqrt{p_{lk}} s_k + \sum_{n\neq k} \sum_{b\in\mathcal{T}_n\cap\mathcal{C}_k} \mathbf{h}_{bk}^H \mathbf{w}_{bn} \sqrt{p_{bn}} s_n \\ & + \sum_{n\neq k} \sum_{d\in\mathcal{T}_n\setminus\mathcal{C}_k} \mathbf{h}_{dk}^H \mathbf{w}_{dn} \sqrt{p_{dn}} s_n + z_k, \forall k \in \mathcal{K}. \end{aligned} \tag{3}$$

In (3), s_k represents the data symbol for MU k with unit power and s_k is independent with $s_n, \forall n \neq k$. In addition, $\mathbf{h}_{lk} \in \mathbb{C}^{N_l \times 1}$ is the channel vector from RRH l to MU k, $\mathbf{w}_{lk} \in \mathbb{C}^{N_l \times 1}$ is the unit-norm precoding vector for MU k at RRH l, p_{lk} is the transmit power allocated by RRH l to MU k, and $z_k \sim \mathcal{CN}(0, \sigma_k^2)$ is the additive white Gaussian noise. It is noted that the first term in the right hand of (3) is the desired signals for MU k, the second term is the intra-cluster interference caused by signals transmitted by RRHs in MU k's cluster to other MUs, and the third term is the inter-cluster interference derived from signals transmitted by RRHs beyond MU k's cluster to other MUs.

Instead of beamforming optimization schemes, we adopt ZF precoding in this work to strike a balance between system performance and implementation complexity. As stated in [23], ZF precoding can completely cancel the intra-cluster interference in (3). Besides, the combination of hybrid clustering and ZF precoding can take full advantage of DoF and fronthaul capacity so as to suppress interference and improve system performance as much as possible. Then, the detailed principles of ZF precoding are given. If RRH l only transmits desired signal to MU k and does not coordinate any MUs, $\mathbf{w}_{lk} = \frac{\mathbf{h}_{lk}}{||\mathbf{h}_{lk}||}$ is designed for MU k. Otherwise, $\mathbf{w}_{lk}$ is skillfully obtained by the method of orthogonal projection [27], [28]. In particular,

$$\mathbf{w}_{lk} = \frac{\mathbf{P}_{l,-k}\mathbf{h}_{lk}}{||\mathbf{P}_{l,-k}\mathbf{h}_{lk}||}, \forall k \in \mathcal{K}, \forall l \in \mathcal{T}_k,$$

where $\mathbf{P}_{l,-k} = \mathbf{I} - \mathbf{H}_{l,-k}(\mathbf{H}_{l,-k}^H \mathbf{H}_{l,-k})^{-1}\mathbf{H}_{l,-k}^H$ denotes the orthogonal projector of $\mathbf{h}_{lk}$ onto the orthogonal complementary subspace of subspace spanned with channel coefficient vectors $\mathbf{h}_{ln}, \forall n \in \mathcal{Q}_l \backslash k$. Moreover, $\mathcal{Q}_l$ represents the set of all served and coordinated MUs by RRH l and $\mathbf{H}_{l,-k}$ is the channel coefficient matrix, whose columns are constructed by $\mathbf{h}_{ln}, \forall n \in \mathcal{Q}_l \backslash k$.

To completely cancel intra-cluster interference with ZF precoding, the overall number of MUs served and coordinated simultaneously by RRH l is restricted by the number of its antennas as follows:

$$\sum_{k\in\mathcal{K}} (t_{lk} + a_{lk}) \leq N_l, \forall l \in \mathcal{L}. \tag{4}$$

Meanwhile, it is not possible for RRH l to establish transmission link and coordination link for MU k at the same time. Hence,

$$t_{lk} + a_{lk} \leq 1, \forall k \in \mathcal{K}, \forall l \in \mathcal{L}. \tag{5}$$

3) Power Allocation: Since intra-cluster interference is cancelled by ZF precoding, the signal-to-interference-plus-noise ratio (SINR) of MU k is given by

$$\gamma_k = \frac{|\sum_{l\in\mathcal{T}_k} \sqrt{p_{lk}} \mathbf{h}_{lk}^H \mathbf{w}_{lk}|^2}{\sum_{n\neq k} |\sum_{d\in\mathcal{T}_n\setminus\mathcal{C}_k} \sqrt{p_{dn}} \mathbf{h}_{dk}^H \mathbf{w}_{dn}|^2 + \sigma_k^2}, \forall k \in \mathcal{K}, \tag{6}$$

and the achievable rate of MU k is expressed as

$$c_k = B_k \log(1 + \gamma_k), \forall k \in \mathcal{K}, \tag{7}$$

where B_k is the bandwidth for MU k. It is observed from (6) and (7) that c_k is jointly determined by hybrid clustering, i.e., $\mathcal{T}$ and $\mathcal{A}$, and transmit power allocation, i.e., $p_{lk}, \forall k \in \mathcal{K}, \forall l \in \mathcal{L}$.

However, clustering and transmit power allocation are always executed in different time-scales [29]. Transmit power allocation is adopted to combat the small scale channel fading varying frequently, which is adjusted in millisecond, while clustering is updated in larger time-scale. Therefore, we apply a simple power allocation policy and only optimize the clustering instead of the joint optimization on clustering and power allocation. Specifically, equal power allocation at each RRH is adopted [30], where all MUs served by the same RRH are allocated with equal transmit power as follows:

$$p_{lk} = \begin{cases} \frac{P_l^{\max}}{S_l}, & \forall k \in \mathcal{K}, \forall l \in \mathcal{T}_k, \\ 0, & \forall k \in \mathcal{K}, \forall l \notin \mathcal{T}_k, \end{cases} \tag{8}$$

where $P_l^{\max}$ is the maximum power budget of RRH l. Note that the transmit power consumption of MU k is zero at any RRH that does not transmit desired signal to MU k.

C. QoS Metric for MUs

To jointly investigate computation provisioning and hybrid clustering, we define a systematic QoS metric for MUs. Specifically, we consider a dynamic system, where the packet arrival process for MU k is Poisson with rate λ_k and packet

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JSAC.2016.2624119, IEEE Journal on Selected Areas in Communications

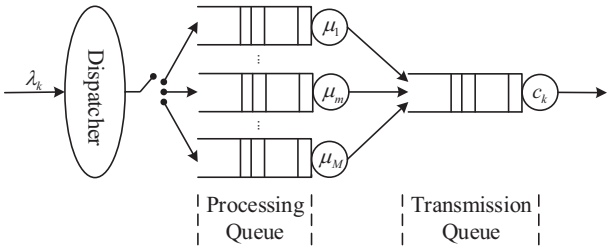

Fig. 2. The tandem queue model for MU k.

sizes for MU k are identical. At first, packets for MU k arrive at a dispatcher and then are routed to its assigned VM for baseband processing. After baseband processing, the processed packets for MU k are shared among its transmitted RRHs, which lie in $\mathcal{T}_k$, via the corresponding fronthauls. At last, RRHs in $\mathcal{T}_k$ apply ZF precoding and allocated transmit power to send the processed packets to MU k while meeting its QoS requirement. Here, quasi-static wireless fading channel conditions are assumed such that, the channel state does not change over the time interval of interest [31].

In summary, the dynamic downlink transmission for MU k can be graphically abstracted as a tandem queue model comprised of processing queue and transmission queue, as shown in Fig. 2. In particular, the processing queue and transmission queue are used to describe baseband processing in the BBU pool and cooperative transmission at RRHs respectively. Moreover, we assume that the service time at VM m follows an exponential distribution with mean $\frac{1}{\mu_m}$. Therefore, the service time of each packet for MU k at its assigned VM follows an exponential distribution with mean $\frac{1}{\nu_k}$, where $\nu_k = \sum_{m\in\mathcal{M}} x_{mk}\mu_m$. Then, the processing queue for MU k is perceived as an M/M/1 queue and the baseband processing delay for MU k is given by

$$\tau_k^{\mathrm{p}} = \frac{1}{\nu_k - \lambda_k}, \forall k \in \mathcal{K}.$$

Subsequently, the processed packets for MU k depart the processing queue and arrive to the transmission queue through the corresponding fronthauls with negligible transport delay [32]. According to the Burke's Theorem [33], the arrival process of transmission queue for MU k (i.e., the departure process of processing queue for MU k) is still Poisson with rate λ_k. Under the assumption of quasi-static wireless channel, all transmitted RRHs in $\mathcal{T}_k$, treated as one virtual access node [34], provide deterministic service time with $\frac{1}{c_k}$ for these identical-size packets. Thus, the transmission queue for MU k can be regarded as an M/D/1 queue and the wireless transmission delay for MU k is expressed as

$$\tau_k^{\mathrm{w}} = \frac{2c_k - \lambda_k}{2c_k(c_k - \lambda_k)}, \forall k \in \mathcal{K}.$$

Therefore, the delay for MU k is jointly determined by τ_k^{p} and τ_k^{w}. Then, the following delay constraint for MU k holds:

$$\frac{1}{\nu_k - \lambda_k} + \frac{2c_k - \lambda_k}{2c_k(c_k - \lambda_k)} \leq \tau_k, \forall k \in \mathcal{K}, \tag{9}$$

where τ_k is the QoS requirement for MU k. In addition,

$$\nu_k > \lambda_k, c_k > \lambda_k, \forall k \in \mathcal{K}, \tag{10}$$

should be satisfied to guarantee the stability of the queuing system [33].

III. Problem Formulation

In this section, we formulate the system power minimization problem constrained by limited fronthauls. Particularly, baseband processing power consumption at VM m, determined by VM assignment, is expressed as $\sum_{k\in\mathcal{K}} x_{mk}\phi_m$. While the power consumption at RRH l, depending on hybrid clustering, is denoted by $\sum_{k\in\mathcal{K}} p_{lk} + ||\sum_{k\in\mathcal{K}} t_{lk}||_0 P_l^{\mathrm{c}}$, where P_l^{c} represents the static power consumption at active RRH l. Then, we denote the system power consumption as

$$\begin{aligned} E(\mathcal{T},\mathcal{A},\mathbf{x}) \quad = \quad & \sum_{m\in\mathcal{M}}\sum_{k\in\mathcal{K}} x_{mk}\phi_m \\ & + \eta \sum_{l\in\mathcal{L}} \left\{ \sum_{k\in\mathcal{K}} p_{lk} + ||\sum_{k\in\mathcal{K}} t_{lk}||_0 P_l^{\mathrm{c}} \right\} \end{aligned} \tag{11}$$

where $\mathbf{x} = [x_{11}, ..., x_{1K}, ..., x_{M1}, ..., x_{MK}]^T$ and $\eta > 0$ is introduced as a factor to strike a balance between these two parts of power consumption. Note that fronthaul power consumption is always modeled as a step function with two levels, which depends on whether the fronthaul is active or not [10]. Thus, (11) is general enough to account the fronthaul power consumption in P_l^{c}.

By jointly optimizing hybrid clustering and computation provisioning, the system power minimization problem can be formulated as

$$\begin{aligned} \mathscr{P}: \min_{\mathcal{T},\mathcal{A},\mathbf{x}} \; & E(\mathcal{T},\mathcal{A},\mathbf{x}) \\ \text{s.t. C1}: \; & \sum_{m\in\mathcal{M}} x_{mk} = 1, \forall k \in \mathcal{K} \\ \text{C2}: \; & \sum_{k\in\mathcal{K}} x_{mk} \leq 1, \forall m \in \mathcal{M} \\ \text{C3}: \; & x_{mk} \in \{0,1\}, \forall k \in \mathcal{K}, \forall m \in \mathcal{M} \\ \text{C4}: \; & \sum_{k\in\mathcal{K}} t_{lk} \leq S_l, \forall l \in \mathcal{L} \\ \text{C5}: \; & \sum_{k\in\mathcal{K}} (t_{lk} + a_{lk}) \leq N_l, \forall l \in \mathcal{L} \\ \text{C6}: \; & t_{lk} + a_{lk} \leq 1, \forall k \in \mathcal{K}, \forall l \in \mathcal{L} \\ \text{C7}: \; & \frac{1}{\nu_k - \lambda_k} + \frac{2c_k - \lambda_k}{2c_k(c_k - \lambda_k)} \leq \tau_k, \forall k \in \mathcal{K} \\ \text{C8}: \; & \nu_k > \lambda_k, c_k > \lambda_k, \forall k \in \mathcal{K}, \end{aligned}$$

where C1-C3 represent constraints on the VM assignment. C4 reflects the limited fronthaul capacity. C5 is derived from the ZF precoding and hybrid clustering makes C6 hold. To meet the QoS requirement for each MU, we have C7. Moreover, C8 is imposed to guarantee the stability of the queueing system. It is observed from C4, C5, and C7 that the increasing number of MUs can make problem $\mathscr{P}$ infeasible. By this time, admission control in [35]–[37] should be introduced to admit

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JSAC.2016.2624119, IEEE Journal on Selected Areas in Communications

a reasonable number of MUs in order to make problem $\mathscr{P}$ feasible. However, this infeasible case is out of the scope of this paper. Specifically, we assume that problem $\mathscr{P}$ is always feasible for the rest of this paper.

Subsequently, we focus on tackling problem $\mathscr{P}$ to make an optimal decision on VM assignment and hybrid clustering. Since the computation capacity in C-RAN is mitigated into the cloud-based BBU pool, the optimal decision is made in the B-BU pool in a centralized way. The decision on VM assignment gives a guidance to packet routing for MU $k, \forall k \in \mathcal{K}$ among the dispatcher and VMs while the obtained hybrid clustering is used to further design the ZF precoder vector $\mathbf{w}_{lk}$ for MU k at its transmitted RRH $l, \forall l \in \mathcal{T}_k$. Once precoding vectors are determined, the BBU pool transmits MU k's data along with the precoding vector $\mathbf{w}_{lk}$ to its transmitted RRH $l, \forall l \in \mathcal{T}_k$ through the fronthaul link. Then, all these transmitted RRHs of MU k adopt the devised precoding vectors to cooperatively transmit data to MU k.

Problem $\mathscr{P}$ includes two groups of key design parameters: VM assignment parameter $\mathbf{x}$ and hybrid clustering parameters $\mathcal{T}$ and $\mathcal{A}$. However, these two group of parameters are integers and closely coupled in C7, which makes problem $\mathscr{P}$ as an integer non-linear programming problem. To tackle the challenging problem $\mathscr{P}$, we take full use of its special structure and further reformulate it as an equivalent hybrid clustering problem ($\mathscr{P}_{\text{HC}}$), where a series of VM assignment problems ($\mathscr{P}_{\text{MA}}$) are embedded. Hence, solving problem $\mathscr{P}$ is equivalent to solving problem $\mathscr{P}_{\text{HC}}$. In the following, we elaborate problem $\mathscr{P}_{\text{HC}}$ and its embedded problem $\mathscr{P}_{\text{MA}}$, respectively.

A. VM Assignment Problem

When a feasible case of $\mathcal{T}$ and $\mathcal{A}$ is given in problem $\mathscr{P}$, we can get a VM assignment problem, denoted by

$$\begin{aligned}\mathscr{P}_{\text{MA}}: \min_{\mathbf{x}} & \sum_{k\in\mathcal{K}}\sum_{m\in\mathcal{M}} x_{mk}\phi_m \\ \text{s.t. C1}: & \sum_{m\in\mathcal{M}} x_{mk} = 1, \forall k \in \mathcal{K} \\ \text{C2}: & \sum_{k\in\mathcal{K}} x_{mk} \le 1, \forall m \in \mathcal{M} \\ \text{C3}: & x_{mk} \in \{0,1\}, \forall k \in \mathcal{K}, \forall m \in \mathcal{M} \\ \text{C9}: & \sum_{m\in\mathcal{M}} x_{mk}\mu_m \ge \frac{1}{\tau_k'} + \lambda_k, \forall k \in \mathcal{K}.\end{aligned}$$

Here, $\tau_k' = \tau_k - \frac{2c_k-\lambda_k}{2c_k(c_k-\lambda_k)}$ is positive in C9, which is a new constraint derived from C7 and C8. In detail, with given $\mathcal{T}$ and $\mathcal{A}$, we have the following two inequalities from C7 and C8, respectively:

$$\frac{1}{\nu_k - \lambda_k} \le \tau_k', \forall k \in \mathcal{K}, \text{ and} \tag{12}$$

$$\nu_k > \lambda_k, \forall k \in \mathcal{K}. \tag{13}$$

It is observed from (12) and (13) that $\nu_k \ge \frac{1}{\tau_k'} + \lambda_k > \lambda_k$ should be satisfied for any MU k. Hence, we finally get the constraint $\nu_k \ge \frac{1}{\tau_k'} + \lambda_k, \forall k \in \mathcal{K}$, i.e., C9.

Furthermore, to formulate the hybrid clustering problem, we introduce a function with respect to $\mathcal{T}$ and $\mathcal{A}$, which is expressed as $\mathscr{G}(\mathcal{T},\mathcal{A})$. Given $\mathcal{T}$ and $\mathcal{A}$, $\mathscr{G}(\mathcal{T},\mathcal{A})$ can be obtained by solving the resultant problem $\mathscr{P}_{\text{MA}}$. Specifically, if the resultant problem $\mathscr{P}_{\text{MA}}$ is feasible, $\mathscr{G}(\mathcal{T},\mathcal{A})$ is its optimal value. Otherwise, $\mathscr{G}(\mathcal{T},\mathcal{A}) = +\infty$.

B. Hybrid Clustering Problem

In this subsection, we formulate the hybrid clustering problem, which aims to achieve the optimal $\mathcal{T}$ and $\mathcal{A}$. By primal decomposition on problem $\mathscr{P}$, the hybrid clustering problem embedded with VM assignment problems is written as:

$$\begin{aligned}\mathscr{P}_{\text{HC}}: \min_{\mathcal{T},\mathcal{A}} & \mathscr{G}(\mathcal{T},\mathcal{A}) + \eta\left\{\sum_{l\in\mathcal{L}} ||\sum_{k\in\mathcal{K}} t_{lk}||_0 P_l^{\text{c}} + \sum_{k\in\mathcal{K}}\sum_{l\in\mathcal{T}_k} p_{lk}\right\} \\ \text{s.t. C4}: & \sum_{k\in\mathcal{K}} t_{lk} \le S_l, \forall l \in \mathcal{L} \\ \text{C5}: & \sum_{k\in\mathcal{K}} (t_{lk} + a_{lk}) \le N_l, \forall l \in \mathcal{L} \\ \text{C6}: & t_{lk} + a_{lk} \le 1, \forall k \in \mathcal{K}, \forall l \in \mathcal{L} \\ \text{C10}: & c_k > \frac{1}{2}(\lambda_k + \frac{1}{\tau_k} + \sqrt{\lambda_k^2 + \frac{1}{\tau_k^2}}), \forall k \in \mathcal{K}.\end{aligned}$$

In problem $\mathscr{P}_{\text{HC}}$, C10 gives the feasible region of transmission rate for each MU, which is deduced from C7 and C8. Firstly, (14) and (15) should be satisfied, respectively.

$$\frac{2c_k - \lambda_k}{2c_k(c_k - \lambda_k)} < \tau_k, \forall k \in \mathcal{K}, \text{ and} \tag{14}$$

$$c_k > \lambda_k, \forall k \in \mathcal{K}. \tag{15}$$

Then, by solving the quadratic inequality (14), we get the following inequality constraint on c_k: $c_k > \frac{1}{2}(\lambda_k + \frac{1}{\tau_k} + \sqrt{\lambda_k^2 + \frac{1}{\tau_k^2}}) \bigcup c_k < \frac{1}{2}(\lambda_k + \frac{1}{\tau_k} - \sqrt{\lambda_k^2 + \frac{1}{\tau_k^2}})$. Since $\frac{1}{2}(\lambda_k + \frac{1}{\tau_k} + \sqrt{\lambda_k^2 + \frac{1}{\tau_k^2}}) > \lambda_k > \frac{1}{2}(\lambda_k + \frac{1}{\tau_k} - \sqrt{\lambda_k^2 + \frac{1}{\tau_k^2}})$ with $\tau_k > 0, \lambda_k > 0$ invariably holds for MU $k \in \mathcal{K}$, the feasible region of c_k is naturally shrunk to $c_k > \frac{1}{2}(\lambda_k + \frac{1}{\tau_k} + \sqrt{\lambda_k^2 + \frac{1}{\tau_k^2}}), \forall k \in \mathcal{K}$.

IV. Optimal Solution

In this section, we present the optimal solution of problem $\mathscr{P}$ by solving the equivalent problem $\mathscr{P}_{\text{HC}}$. Particularly, the embedded problem $\mathscr{P}_{\text{MA}}$ is tackled to obtain the optimal $\mathbf{x}$ from the viewpoint of bipartite graph matching and the optimal $\mathcal{T}$ and $\mathcal{A}$ is achieved by using exhaustive search method.

A. Exhaustive Search Method for Problem $\mathscr{P}_{HC}$

It is observed from C6 in problem $\mathscr{P}_{\text{HC}}$ that (t_{lk}, a_{lk}) can be (0,1), (1,0), or (0,0). Thus, there are totally 3^{KL} possible combinations on $\mathcal{T}$ and $\mathcal{A}$. However, the given $\mathcal{T}$ and $\mathcal{A}$ are not always feasible for problem $\mathscr{P}_{\text{HC}}$. This motivates us to define the feasibility condition of $\mathcal{T}$ and $\mathcal{A}$.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JSAC.2016.2624119, IEEE Journal on Selected Areas in Communications

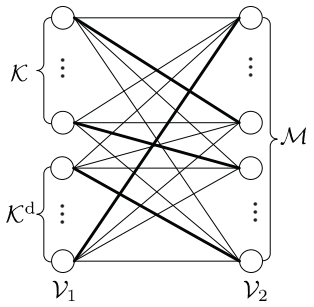

Fig. 3. An illustration of $\mathcal{G}(\mathcal{V}_1 \bigcup \mathcal{V}_2, \mathcal{E})$, where the weight of each edge is given by (16). The thick lines are the optimal matching between $\mathcal{V}_1$ and $\mathcal{V}_2$.

Definition 1. The given $\mathcal{T}$ and $\mathcal{A}$ are feasible when the following two conditions are both satisfied:

1) make C4-C6 and C10 hold;
2) make the resultant problem $\mathscr{P}_{\mathrm{MA}}$ feasible, i.e., $\mathscr{G}(\mathcal{T}, \mathcal{A}) \neq +\infty$.

By adopting the exhaustive search method, we first check the feasibility of each possible $\mathcal{T}$ and $\mathcal{A}$ and then obtain the optimal $\mathcal{T}$ and $\mathcal{A}$ from all the feasible solutions as the one that achieves the minimum system power consumption.

B. Matching Approach for Problem $\mathscr{P}_{MA}$

When the given $\mathcal{T}$ and $\mathcal{A}$ make C4-C6 and C10 hold, we further solve a problem $\mathscr{P}_{\mathrm{MA}}$ to obtain $\mathscr{G}(\mathcal{T}, \mathcal{A})$. Fortunately, we find that solving problem $\mathscr{P}_{\mathrm{MA}}$ can be regarded as a matching procedure between MUs and VMs. Therefore, we apply a matching approach in graph theory to solve problem $\mathscr{P}_{\mathrm{MA}}$ exactly and effectively. For more details, by exploiting the structure of problem $\mathscr{P}_{\mathrm{MA}}$, a perfect bipartite graph $\mathcal{G}(\mathcal{V}_1 \bigcup \mathcal{V}_2, \mathcal{E})$ can be constructed by adding $M - K$ dummy vertices (constituting set $\mathcal{K}^{\mathrm{d}}$) in vertex set $\mathcal{V}_1$. Vertex set $\mathcal{V}_1 = \mathcal{K} \bigcup \mathcal{K}^{\mathrm{d}}$ and vertex set $\mathcal{V}_2 = \mathcal{M}$ in $\mathcal{G}(\mathcal{V}_1 \bigcup \mathcal{V}_2, \mathcal{E})$ are completely disjunct, as shown in Fig. 3. $\mathcal{E} = \{(k, m), \forall k \in \mathcal{V}_1, \forall m \in \mathcal{V}_2\}$ represents the set of edges between $\mathcal{V}_1$ and $\mathcal{V}_2$. By designing weight value $b(k, m)$ on edge $(k, m) \in \mathcal{E}, \forall k \in \mathcal{V}_1, \forall m \in \mathcal{V}_2$, problem $\mathscr{P}_{\mathrm{MA}}$ can be transformed as a minimum weight perfect matching (MWPM) problem.

Let $\mathcal{M}_k = \{m | \mu_m \geq \frac{1}{\tau_k^r} + \lambda_k, m \in \mathcal{M}\}$ represent the set of feasible VMs for MU $k \in \mathcal{K}$, which is derived from C9 in problem $\mathscr{P}_{\mathrm{MA}}$. Hence, problem $\mathscr{P}_{\mathrm{MA}}$ is infeasible when MU $k \in \mathcal{K}$ is provisioned with VM $m \notin \mathcal{M}_k$. Correspondingly, $b(k, m) = +\infty$ is allocated on edge $(k, m), \forall k \in \mathcal{K}, \forall m \notin \mathcal{M}_k$. When VM $m \in \mathcal{M}_k$ is assigned to MU $k \in \mathcal{K}$, MU $k \in \mathcal{K}$ consumes the processing power ϕ_m and thus, $b(k, m) = \phi_m, \forall k \in \mathcal{K}, \forall m \in \mathcal{M}_k$. Moreover, the dummy vertices cannot produce additional power consumption in the objective value of problem $\mathscr{P}_{\mathrm{MA}}$, i.e., $b(k, m) = 0, \forall k \in \mathcal{K}^{\mathrm{d}}, \forall m \in \mathcal{M}$. In summary, we have

$$b(k, m) = \begin{cases} +\infty, & \forall k \in \mathcal{K}, \forall m \notin \mathcal{M}_k \\ \phi_m, & \forall k \in \mathcal{K}, \forall m \in \mathcal{M}_k \\ 0, & \forall k \in \mathcal{K}^{\mathrm{d}}, \forall m \in \mathcal{M}. \end{cases} \tag{16}$$

Hence, by transforming C9 and objective function in problem $\mathscr{P}_{\mathrm{MA}}$ into the weight design of edges in $\mathcal{G}(\mathcal{V}_1 \bigcup \mathcal{V}_2, \mathcal{E})$, we can reformulate problem $\mathscr{P}_{\mathrm{MA}}$ as a standard MWPM problem as follows:

$$\mathscr{P}_{\mathrm{MA}}^{\mathrm{MWPM}} : \min_{\mathbf{x}} \sum_{k \in \mathcal{K}} \sum_{m \in \mathcal{M}} b(k, m) x_{mk}$$
$$\text{s.t. C1-C3,}$$

which can be effectively solved by the Hungarian method with computational complexity $\mathcal{O}(M^3)$ [38]. If the optimal value of problem $\mathscr{P}_{\mathrm{MA}}^{\mathrm{MWPM}}$ is $+\infty$, i.e., at least one MU is assigned with an infeasible VM, problem $\mathscr{P}_{\mathrm{MA}}$ is infeasible. Otherwise, problem $\mathscr{P}_{\mathrm{MA}}$ is feasible and its optimal solution can be directly obtained by solving problem $\mathscr{P}_{\mathrm{MA}}^{\mathrm{MWPM}}$.

C. Complexity Analysis

We call the aforementioned exhaustive search method for system power minimization as ExhSearch. Although the ExhSearch can make the optimal decision on hybrid clustering and VM assignment, it is evident that this method has exponential complexity. In detail, there are totally 3^{KL} possible cases for $\mathcal{T}$ and $\mathcal{A}$ and for each given $\mathcal{T}$ and $\mathcal{A}$, an embedded problem $\mathscr{P}_{\mathrm{MA}}$ may be solved with complexity $\mathcal{O}(M^3)$. Hence, the computational complexity of ExhSearch is approximate to $\mathcal{O}(3^{KL} M^3)$.

V. A Low Complexity Algorithm For The Hybrid Clustering Problem

As described in the last section, the ExhSearch has prohibitive complexity with increasing number of RRHs and MUs, which motivates us to devise a low complexity algorithm. In this section, we propose a greedy algorithm, referred to as GreHybrid, to solve the hybrid clustering problem with low complexity. The GreHybrid is summarized in Algorithm 1, which mainly consists of two procedures as follows:

- Initialization procedure (lines 1-3): In this procedure, we aim to find the $\mathcal{T}$ and $\mathcal{A}$, which can meet all the constraints in problem $\mathscr{P}_{\mathrm{HC}}$, i.e., C4-C6 and C10. Firstly, an association rule, which jointly considers channel gains of MUs and available resources at RRHs, is devised to identify initial $\mathcal{T}$ and $\mathcal{A}$ while fulfilling C4-C6. Then, an appropriate link is successively appended to one MU to update initial $\mathcal{T}$ and $\mathcal{A}$ until C4-C6 and C10 are all satisfied.
- Iteration procedure (lines 7-11): Based on the $\mathcal{T}$ and $\mathcal{A}$ derived from the initialization procedure, we then update them step by step to reduce the system power consumption. In the iteration procedure, the best link is greedily selected from the set of unused links and is appended to an appropriate MU to reduce the system power consumption in each iteration. This procedure is continually executed until the system power consumption stops decreasing.

Subsequently, we elaborate these two procedures in the following two subsections.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JSAC.2016.2624119, IEEE Journal on Selected Areas in Communications

Algorithm 1 Greedy algorithm for the hybrid clustering problem (GreHybrid)

Initialize $\mathcal{A} = \mathbf{0}$, $N^{\text{rem}} = [N_1, ..., N_L]$, and $S^{\text{rem}} = [S_1, ..., S_L]$;
Identify the initial $\mathcal{T}$ based on the devised association rule to meet C4-C6 and update N^{rem} and S^{rem};
Update the initial $\mathcal{T}$ and $\mathcal{A}$, N^{rem}, and S^{rem} to improve $c_k, \forall k \in \mathcal{K}$ in order to meet C10;
if N^{rem}=[0,...,0] **then**
  Problem $\mathscr{P}$ is infeasible and algorithm terminates.
else
  repeat
    Obtain the latest $\mathcal{T}$, $\mathcal{A}$, N^{rem}, and S^{rem};
    Based on (19), assign the best link $(\mathcal{R}^*_{k^*}, \mathcal{Y}^*_{k^*})$ to MU k^* and solve the resultant problem $\mathscr{P}_{\text{MA}}$;
    Calculate the amount of system power reduction $\theta(\mathcal{R}^*_{k^*}, \mathcal{Y}^*_{k^*})$;
  until $\theta(\mathcal{R}^*_{k^*}, \mathcal{Y}^*_{k^*}) < 0$ or $N^{\text{rem}} = [0, ..., 0]$
  if N^{rem}=[0,...,0] and $\mathscr{G}(\mathcal{T}, \mathcal{A}) = +\infty$ **then**
    Problem $\mathscr{P}$ is infeasible and algorithm terminates.
  end if
  Output the final $\mathcal{T}$, $\mathcal{A}$, and $\mathbf{x}$.
end if

A. Initialization Procedure

The initialization procedure is proposed to find the $\mathcal{T}$ and $\mathcal{A}$, which can make C4-C6 and C10 satisfied. C4-C6 are regarded as resource constraints and thus, we introduce $N^{\text{rem}} = [n_1, ..., n_L]$ and $S^{\text{rem}} = [s_1, ..., s_L]$ to denote the remaining antenna and fronthaul resources at RRHs. At first, we initialize $\mathcal{A} = \mathbf{0}$, $N^{\text{rem}} = [N_1, ..., N_L]$, and $S^{\text{rem}} = [S_1, ..., S_L]$. Then, the following two steps are executed successively to fulfill C4-C6 and C10:

1) Identify the initial $\mathcal{T}$ based on the devised association rule: Firstly, any MU k is associated with its best RRH l_k according to the largest channel gain policy:

$$l_k = \arg\max_{l \in \mathcal{L}} ||h_{lk}||^2, \forall k \in \mathcal{K}. \tag{17}$$

Then, we set $t_{l_k k} = 1, \forall k \in \mathcal{K}$ to establish the transmission link between RRH l_k and MU k and set other elements in $\mathcal{T}$ as zeros at the same time. Furthermore, the remaining antenna and fronthaul resources at any RRH $l \in \mathcal{L}$ are calculated as $n_l = n_l - \sum_{k \in \mathcal{K}} t_{lk}$ and $s_l = s_l - \sum_{k \in \mathcal{K}} t_{lk}$, respectively.

The $\mathcal{T}$ derived from the largest channel gain policy may incur C4 and C5 becoming infeasible. Specifically, if $n_l < 0$ or $s_l < 0$, antenna or fronthaul resources at RRH l are over-utilized. Hence, a further pruning is imperative. We introduce $\mathcal{L}^{\text{over}} = \{l | n_l < 0 \bigcup s_l < 0, l \in \mathcal{L}\}$ to indicate the set of over-utilized RRHs and $\mathcal{E}_l$ to stand for the set of removed MUs from RRH l. For each RRH $l \in \mathcal{L}^{\text{over}}$, its associated MUs is sorted in descending order in terms of channel gains and then, the last $|\mathcal{E}_l|$ MUs are all removed from RRH $l \in \mathcal{L}^{\text{over}}$ to satisfy C4 and C5. Meanwhile, the corresponding elements in $\mathcal{T}$ are set as zeros. Note that the remaining resources in RRH l have a direct impact on the number of removed MUs from RRH $l \in \mathcal{L}^{\text{over}}$, i.e., $|\mathcal{E}_l|$. In detail,

$$|\mathcal{E}_l| = \begin{cases} \max\{|n_l|, |s_l|\}, & n_l < 0, s_l < 0 \\ |n_l|, & n_l < 0, s_l \geq 0 \\ |s_l|, & n_l \geq 0, s_l < 0. \end{cases}$$

Subsequently, any MU $k \in \bigcup_{l \in \mathcal{L}^{\text{over}}} \mathcal{E}_l$ prefers to access RRH $b \in \mathcal{L} \backslash \mathcal{L}^{\text{over}}$, which can provide the largest channel gain to MU k among all the under-utilized RRHs. However, RRH $b \in \mathcal{L} \backslash \mathcal{L}^{\text{over}}$ can only accommodate at most $\min\{n_b, s_b\}$ MUs due to limited available resources. Thus, RRH b gives preference to MUs with larger channel gains and rejects the other MUs. If MU $k \in \bigcup_{l \in \mathcal{L}^{\text{over}}} \mathcal{E}_l$ is accepted by RRH b, $t_{bk} = 1$. Otherwise, MU k is added to the set $\mathcal{E}_b$. When $|\mathcal{E}_b| > 0$, RRH b uses up its fronthaul or antenna resources and thus is added to $\mathcal{L}^{\text{over}}$. Once this access procedure is finished, any RRH $l \in \mathcal{L}$ updates its remaining resources in terms of $n_l = n_l - \sum_{k \in \mathcal{K}} t_{lk}$ and $s_l = s_l - \sum_{k \in \mathcal{K}} t_{lk}$. Then, such access procedure is executed repeatedly until $\bigcup_{l \in \mathcal{L}^{\text{over}}} \mathcal{E}_l = \emptyset$.

Based on this devised association rule, each RRH is not over-utilized and each MU is served by one transmitted RRHs and thus, C4-C6 are satisfied preliminarily.

2) Update the initial $\mathcal{T}$ and $\mathcal{A}$ to improve $c_k, \forall k \in \mathcal{K}$: Once C4-C6 hold, our next concern is on C10. We introduce $\mathcal{K}^{\text{out}} = \{k | c_k \leq \frac{1}{2}(\lambda_k + \frac{1}{\tau_k} + \sqrt{\lambda_k^2 + \frac{1}{\tau_k^2}})\}$ to denote the set of MUs, whose transmission rate cannot satisfy C10. If $\mathcal{K}^{\text{out}} = \emptyset$, the targeted $\mathcal{T}$ and $\mathcal{A}$ are found. Otherwise, we find out MU k' with the smallest transmission rate, i.e., $k' = \arg\min_{k \in \mathcal{K}^{\text{out}}} c_k$ and then, successively select the best link from the unused links and append it to MU k' to improve c'_k. Here, two matrices $\mathcal{R}_k = [r_{lk}]_{L \times K}$ and $\mathcal{Y}_k = [y_{lk}]_{L \times K}$ are introduced, in which only one element is 1 and all other elements are 0. The element "1" only arises in the k-th column in $\mathcal{R}_k$ or $\mathcal{Y}_k$ to indicate a transmission link or a cooperative link for MU k. For instance, $r_{lk} = 1$ represents that the selected link for MU k is the transmission link between RRH l and MU k. Likewise, $y_{lk} = 1$ means that the cooperation link between RRH l and MU k is selected for MU k.

Next, we define the feasible link set $\mathcal{F}_k = \{(\mathcal{R}_k, \mathcal{Y}_k)\}$ for MU k, in which $\mathcal{R}_k + \mathcal{T}$ and $\mathcal{Y}_k + \mathcal{A}$ should meet C4-C6 as well. Specifically, $r_{lk} = 1$ or $y_{lk} = 1$ when $t_{lk} = 0$ and $a_{lk} = 0$, $\forall l \in \{l | n_l > 0, s_l > 0, l \in \mathcal{L}\}$ or $y_{lk} = 1$ when $t_{lk} = 0$ and $a_{lk} = 0$, $\forall l \in \{l | n_l > 0, s_l = 0, l \in \mathcal{L}\}$. All the possible $(\mathcal{R}_k, \mathcal{Y}_k)$ construct the feasible link set $\mathcal{F}_k$ for MU k. Then, the best link for MU k' is selected according to the following policy:

$$(\mathcal{R}^*_{k'}, \mathcal{Y}^*_{k'}) = \arg \max_{(\mathcal{R}_{k'}, \mathcal{Y}_{k'}) \in \mathcal{F}_{k'}} \delta(\mathcal{R}_{k'}, \mathcal{Y}_{k'}), \tag{18}$$

where $\delta(\mathcal{R}_{k'}, \mathcal{Y}_{k'}) = c_{k'}(\mathcal{T} + \mathcal{R}_{k'}, \mathcal{A} + \mathcal{Y}_{k'}) - c_{k'}(\mathcal{T}, \mathcal{A}) - \sum_{n \neq k'}(c_n(\mathcal{T}, \mathcal{A}) - c_n(\mathcal{T} + \mathcal{R}_{k'}, \mathcal{A} + \mathcal{Y}_{k'}))$ represents the difference between the increment of c'_k and the total decrement of all c_n, $\forall n \neq k'$. This proposed policy emphasizes that the selected link for MU k' should increase c'_k as much as possible while decreasing $c'_n, \forall n \neq k'$ to the least extent. After appending the selected link to MU k', the current $\mathcal{T}$, $\mathcal{A}$, N^{rem}, S^{rem}, and $\mathcal{K}^{\text{out}}$ are updated accordingly. This step is iteratively executed until $\mathcal{K}^{\text{out}} = \emptyset$ or $N^{\text{rem}} = [0, ..., 0]$.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JSAC.2016.2624119, IEEE Journal on Selected Areas in Communications

If this step terminates at $N^{\text{rem}} = [0, ..., 0]$, we cannot improve MUs' transmission rates to adhere to C10 and Algorithm 1 terminates. Otherwise, we obtain the targeted $\mathcal{T}$ and $\mathcal{A}$, which make all the constraints in problem $\mathscr{P}_{\text{HC}}$ hold. After the initialization procedure, the first feasibility condition of $\mathcal{T}$ and $\mathcal{A}$ can be guaranteed in terms of Definition 1. Thus, to obtain the solution for problem $\mathscr{P}_{\text{HC}}$, we should further update $\mathcal{T}$ and $\mathcal{A}$ using the remaining resources at RRHs to meet the second feasibility condition and reduce the system power consumption at the same time. This observation motivates us to devise the iteration procedure.

B. Iteration Procedure

In the iteration procedure, we introduce $\theta(\mathcal{R}_k, \mathcal{Y}_k) = \mathscr{H}(\mathcal{T}, \mathcal{A}) - \mathscr{H}(\mathcal{T} + \mathcal{R}_k, \mathcal{A} + \mathcal{Y}_k)$ to stand for the amount of system power reduction when link $(\mathcal{R}_k, \mathcal{Y}_k)$ is added to the current $(\mathcal{T}, \mathcal{A})$. Therein, $\mathscr{H}(\mathcal{T}, \mathcal{A})$ is the system power consumption with given $\mathcal{T}$ and $\mathcal{A}$, which consists of power consumption at RRHs, i.e., $\eta\{\sum_{l\in\mathcal{L}} || \sum_{k\in\mathcal{K}} t_{lk}||_0 P_l^{\text{c}} + \sum_{k\in\mathcal{K}} \sum_{l\in\mathcal{T}_k} p_{lk}\}$, and power consumption in the BBU pool, i.e., $\mathscr{G}(\mathcal{T}, \mathcal{A})$. Likewise, $\mathscr{H}(\mathcal{T} + \mathcal{R}_k, \mathcal{A} + \mathcal{Y}_k)$ is the system power consumption with $\mathcal{T} = \mathcal{T} + \mathcal{R}_k$ and $\mathcal{A} = \mathcal{A} + \mathcal{Y}_k$. Then, we select the best MU-link pair as follows:

$$(k^*, \mathcal{R}_{k*}^*, \mathcal{Y}_{k^*}^*) = \arg \max_{k\in\mathcal{K}, (\mathcal{R}_k, \mathcal{Y}_k)\in\mathcal{F}_k} \theta(\mathcal{R}_k, \mathcal{Y}_k). \quad (19)$$

and add the selected link $(\mathcal{R}_{k*}^*, \mathcal{Y}_{k^*}^*)$ to the best MU k^* in each iteration. Such iteration is executed successively until $\theta(\mathcal{R}_k, \mathcal{Y}_k) < 0$ or $N^{\text{rem}} = [0, ..., 0]$. Note that $\mathcal{T}$, $\mathcal{A}$, N^{rem}, and S^{rem} should be updated in each iteration. If iteration procedure terminates at $N^{\text{rem}} = [0, ..., 0]$ and $\mathscr{G}(\mathcal{T}, \mathcal{A}) = +\infty$ also holds by this time, problem $\mathscr{P}_{\text{HC}}$ is infeasible. This is because, the resultant problem $\mathscr{P}_{\text{MA}}$ is infeasible and no antenna resources can be further used to update $\mathcal{T}$ and $\mathcal{A}$ to make it feasible. Namely, the second feasible condition of $\mathcal{T}$ and $\mathcal{A}$ cannot be satisfied in this case. Otherwise, we can output the final $\mathcal{T}$, $\mathcal{A}$, and x after the iteration procedure.

C. Convergency and Complexity Analysis

1) Convergency: The convergence of GreHybrid can be guaranteed. It is observed from (6) and (7) that an additional transmission/cooperation link for MU k is useful to improve its transmission rate. However, an additional transmission link for MU k may diminish other MUs' transmission rates, while an additional cooperation link for MU k has no effect on other MUs' transmission rates. Hence, we select the best link (transmission or cooperation link) for MU k according to (18) to improve its transmission rate as much as possible while decreasing other MUs' transmission rates to the least extent. On this basis, the transmission rates of all MUs in $\mathcal{K}^{\text{out}}$ can be improved such that we can always find out the $\mathcal{T}$ and $\mathcal{A}$ to satisfy C4-C6 and C10 in the initialization procedure.

Moreover, the tradeoff between the power consumption in the BBU pool and that at RRHs provides insights for system power saving. In particular, we observe from C7 that increasing transmission rates for MUs at the cost of increasing power consumption at RRHs can result in decreasing processing power consumption in the BBU pool, which dominates the system power consumption in C-RAN. Therefore, by adding more links to MUs, iteration procedure in the GreHybrid aims to improve MUs' transmission rates to reduce the system power consumption. Moreover, adding links successively in the iteration procedure guarantees that the feasible transmission rate for problem $\mathscr{P}_{\text{HC}}$ can always be found and feasible region of the resultant problem $\mathscr{P}_{\text{MA}}$ are zoomed in. Hence, feasible solutions for problem $\mathscr{P}_{\text{HC}}$ are always found to reduce the system power consumption. Finally, iteration procedure terminates when the system power consumption stops decreasing or resources are exhausted.

2) Complexity Analysis: The computational complexity of GreHybrid is dominated by the iteration procedure, where a series of problems $\mathscr{P}_{\text{MA}}$ should be solved. Hence, the worst case in the GreHybrid is that the initialization procedure finishes after each MU is associated with the RRH based on (17). By this time, $LK - K$ iterations should be executed in the iteration procedure. Meanwhile, in the i-th iteration, the search space to find the optimal best MU-link pair in (19) is $2(LK - K - i + 1)$ and for each search, at most one problem $\mathscr{P}_{\text{MA}}$ is tackled with complexity $\mathcal{O}(M^3)$. As a result, the computational complexity of GreHybrid is approximate to $\mathcal{O}(\sum_{i=1}^{LK-K} 2(LK-K-i+1)M^3) = \mathcal{O}(K^2L^2M^3)$, which is much lesser than the ExhSearch with complexity $\mathcal{O}(3^{KL}M^3)$. The complexity of GreHybrid monotonously increases with the network scale and thus, we can further exploit the parallel computing environment in the BBU pool to speed up the execution of the iteration procedure. Specifically, by parallelly calculating the amount of system power reduction for all the feasible MU-link pairs, the best MU-link pair in (19) can be quickly picked out in each iteration.

VI. Simulation Results

Extensive simulation results are presented in this section to validate our proposed joint optimization schemes, i.e., the ExhSearch and the GreHybrid. On one hand, we illustrate the advantages of joint optimization in terms of system power consumption and tradeoffs between power consumption in the BBU pool and that at RRHs. On the other hand, the benefits from the hybrid clustering are showed compared with the clustering schemes under CoMP-JP, CoMP-CB, and NonCoMP mode.

A. Simulation Setup

We construct $M = 40$ heterogeneous VMs with computation capacity (in cycle/s) shown in Table I in random way[1] [39]. Then, power consumption for VM m is calculated by $\phi_m = \alpha(\mu_m)^3$ [39], [40], where μ_m is in cycle/s and $\alpha = 10^{-26}$ is set to match the practical measure [41], [42]. Moreover, we assume that 1900 cycles are required to process per byte data in the BBU pool [41], [42]. Hence, we can obtain

[1]Each element in Table I is generated following $\mathcal{CN}(10^9, 10^9)$. Specifically, when the generated computation capacity for one VM is non-positive, it is generated again until positive.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JSAC.2016.2624119, IEEE Journal on Selected Areas in Communications

TABLE I
COMPUTATION CAPACITY ($\times 10^8$ CYCLE/S) FOR $M = 40$ HETEROGENEOUS VMS.

μ_1~μ_8	17.0631	7.2732	18.1929	3.6696	16.8422	19.4844	24.0913	18.1367
μ_9~μ_{16}	8.2180	9.5478	3.6041	5.6170	16.2310	10.8226	11.5066	18.6616
μ_{17}~μ_{24}	15.8198	12.2973	19.8624	10.6600	9.0836	28.7490	7.7067	11.9286
μ_{25}~μ_{32}	5.5359	18.4361	22.0276	9.1203	19.8941	6.6291	4.9407	5.4500
μ_{33}~μ_{40}	24.1365	12.8006	3.4723	3.9433	5.2432	18.7145	10.1334	11.8770

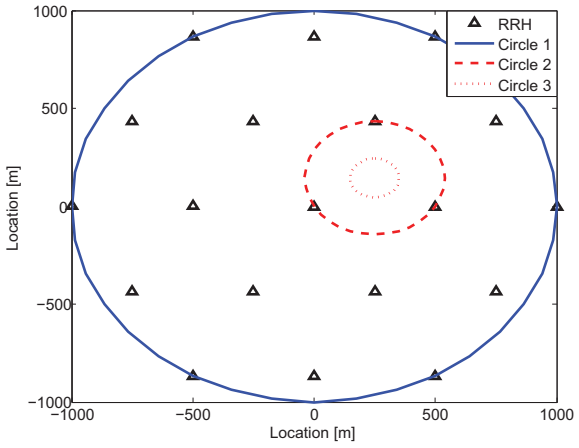

Fig. 4. Simulation scenario.

μ_m bit/s = $(8/1900) \times \mu_m$ cycle/s to calculate the baseband processing delay for MUs.

Moreover, we set a 19-RRH wireless network with 0.5 km inter-RRH distance, as illustrated in Fig. 4. Here, we adopt homogeneous settings for RRHs and MUs, i.e., $N_1 = ... = N_L = N$, $S_1 = ... = S_L = S$, $P_1^{\mathrm{c}} = ... = P_L^{\mathrm{c}} = 2$ W, $P_1^{\max} = ... = P_L^{\max} = 1$ W, $\lambda_1 = ... = \lambda_K = \lambda$ Mbps, and $\tau_1 = ... = \tau_K = \tau$ ms. The distance-dependent path loss model between the RRH and the MU is given by

$$L(\mathrm{dB}) = 128.1 + 37.6 \log_{10}(d),$$

with d in the unit of km [12]. Small scale channel fading is generated with a normalized Rayleigh fading. The power density of noise is -174 dBm/Hz with the channel bandwidth of 10 MHz. Besides, we set $\eta = 1$ if not specified.

B. Advantages of Joint Optimization

Since the complexity of the ExhSearch exponentially increases with the number of RRHs and MUs, we adopt a small-scale wireless network to show the superiorities of the ExhSearch and the GreHybrid and the small performance gap between them in this subsection. Specifically, we consider 3 RRHs, located in Circle 2 of Fig. 4, to control the computation complexity and the simulation time. Meanwhile, the single-antenna MUs are randomly distributed within Circle 3 of 100 m radius. The baseline schemes are listed as follows:

- **Maximum computation provisioning scheme (MCP)**. The MCP focuses on the optimization of hybrid clustering for cooperative transmission ignoring flexible VM assignment. For VM assignment, the MCP aims to reduce processing delay for MUs as much as possible. Specifically, K highest computation capacity VMs are assigned to K MUs by solving an MWPM problem.
- **Static joint processing scheme (StaJP)**. In this scheme, optimal VMs are assigned to MUs in the BBU pool while static clustering under CoMP-JP mode in [13] is adopted for cooperative transmission. This static clustering scheme under the constraint of maximum acceptable number of MUs in each RRH was devised based on the received signal strength.

Figs. 5 and 6 show the effect of the number of MUs on the system power consumption under different λ and τ, where $N = 4$ and $S = 2$. More specific, we set $\tau = 200$ and $\lambda = 1$ for Figs. 5 and 6, respectively. We can find that the system power consumption increases with λ and decreases with τ for all the schemes, including the ExhSearch, the GreHybrid, the StaJP, and the MCP. The larger the number of MUs is, the more system power is consumed with given λ and τ. Moreover, when packet arrival rate is larger or QoS requirement is stricter, lesser number of MUs can be supported simultaneously in the resource-constrained system. Hence, at most 5 MUs can be served with $\lambda = 1.5$ while at most 4 MUs can be served with $\lambda = 2.5$, as shown in Fig. 5. Likewise, we can find from Fig. 6 that the maximum of served MUs is reduced from 5 to 3 with QoS requirement changed from $\tau = 300$ to $\tau = 150$. Note that the ExhSearch has prohibitive complexity with increasing number of MUs and thus, we only simulate for the ExhSearch with the number of MUs not more than 3.

It is observed from Figs. 5 and 6 that the GreHybrid has the close performance as the ExhSearch, which obtains the optimal solution for the original problem $\mathscr{P}$. Besides, the GreHybrid reduces the system power consumption effectively compared with baselines under the same λ and τ. This is because the GreHybrid is devised to dynamically adjust the processing power and transmit power to fit the incoming traffic and QoS requirement. However, the MCP always over-provisions computational resources in the BBU pool and the StaJP always excessively or insufficiently allocates transmit power at RRHs.

Specifically, the MCP provides higher processing speed than that required by the incoming traffic, thereby wasting processing power consumption. Meanwhile, the processing power consumption dominates the system power consumption and thus, the system power consumed by the MCP is the highest. On the other hand, the MCP allocates K VMs with the highest computation capacity to K MUs, which leads to the same processing power consumption under different λ and τ. Hence, the system power consumptions for the MCP under different λ and τ are almost overlap. For the StaJP, the wireless transmission rate is over provisioned when the traffic load is

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JSAC.2016.2624119, IEEE Journal on Selected Areas in Communications

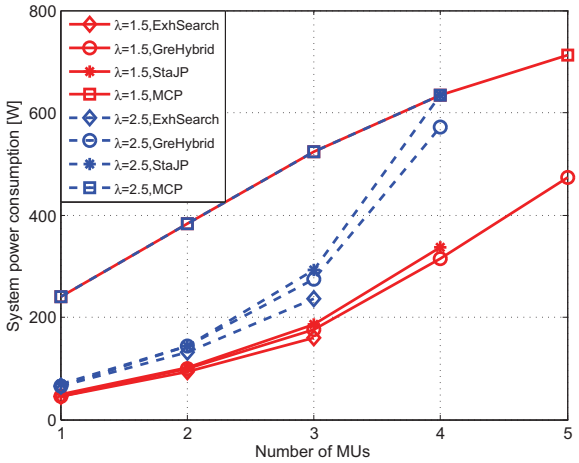

Fig. 5. System power consumption vs. number of MUs under various λ.

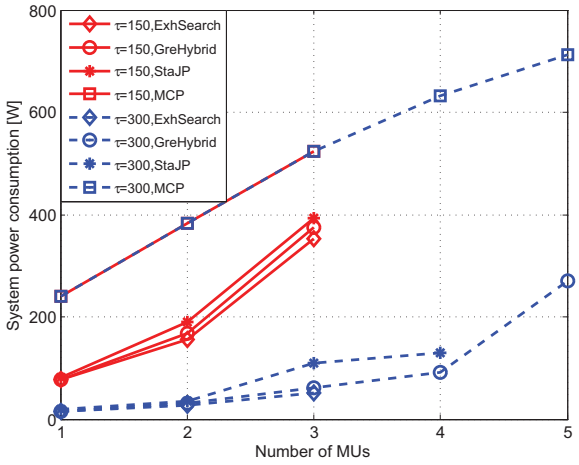

Fig. 6. System power consumption vs. number of MUs under various τ.

small. In contrast, the wireless transmission rate cannot afford higher traffic load for the StaJP and thus, transmission outages occur. In particular, as shown in Figs. 5 and 6, outage occurs for the StaJP with $\lambda = 1.5$ or $\tau = 300$ when the number of MUs is 5.

To further show the advantage of our proposed schemes, Fig. 7 depicts tradeoffs between power consumption in the BBU pool and that at RRHs by leveraging the ExhSearch and the GreHybrid. Here, we set $N = 3$, $K = 3$, $S = 2$, $\lambda = 1$, and $\tau = 200$. It is observed from Fig. 7 that the GreHybrid strikes a little worse balance between these two parts of power consumption compared with the ExhSearch. However, the performance loss of the GreHybrid is acceptable, along with the sharply reduced computational complexity. Moreover, the tradeoff relationship illustrates that we can use more computational resources to exchange for the saving on communication resources. In contrary, computational resources can be saved as well by using more communication resources. By exploiting the tradeoff between communication and computational resources, the system power consumption management in C-RAN becomes more flexible and intelligent.

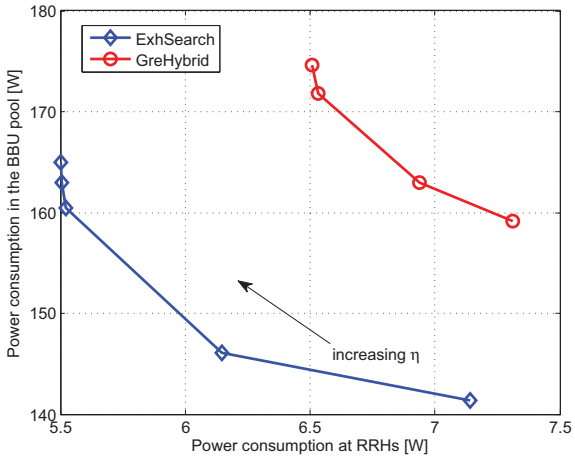

Fig. 7. Tradeoffs between power consumption in the BBU pool and that at RRHs.

C. Advantages of Hybrid Clustering

As stated in the last subsection, the performance of the GreHybrid is in close proximity to that of the ExhSearch. Therefore, in this subsection, only the GreHybrid is simulated to show the superiorities of hybrid clustering under the condition of limited fronthaul capacity. By this time, we consider a large-scale wireless network including 19 RRHs and randomly distribute the signal-antenna MUs within Circle 1 in Fig. 4. The following schemes are treated as baselines:

- **Greedy joint processing (GreJP)**: Compared with the GreHybrid, the GreJP considers the joint processing regardless of coordinated beamforming. Namely, the cooperative matrix $\mathcal{A}$ is set to be all zero and the transmission matrix $\mathcal{T}$ is optimized by adopting Algorithm 1.
- **Greedy coordinated beamforming (GreCB)**: In this scheme, only one RRH transmits desired signal to the MU and other RRHs avoid interfering with the MU. Specifically, the transmission matrix $\mathcal{T}$ is determined according to the first step of initialization procedure in Algorithm 1. In the subsequent procedures, only cooperation links are added to MUs to optimize the cooperative matrix $\mathcal{A}$.
- **NonCoMP**: In the NonCoMP, only one RRH transmits desired signal to the MU. Here, the cooperative matrix $\mathcal{A}$ is set as all zero and the transmission matrix $\mathcal{T}$ is set based on the first step of initialization procedure in Algorithm 1.

It is observed from problem $\mathscr{P}_{\text{HC}}$ that the relationship between the number of antennas N and fronthaul capacity S has a significant effect on the performance of hybrid clustering. Hence, we demonstrate the system power consumption under different S in Fig. 8, where $N = 8$, $K = 15$, $\lambda = 1$, and $\tau = 500$. Since the GreHybrid can fully exploit the antennas resources and fronthaul capacity, it consumes the least system power. On the contrary, the NonCoMP consumes the most system power. Moreover, performances of the GreJP and the GreCB depend on the relationship between N and S. Particularly, when the fronthaul capacity is lacking and antenna resources are sufficient, such as $S = 1$, the GreCB outperforms the GreJP. However, when the fronthaul capacity

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JSAC.2016.2624119, IEEE Journal on Selected Areas in Communications

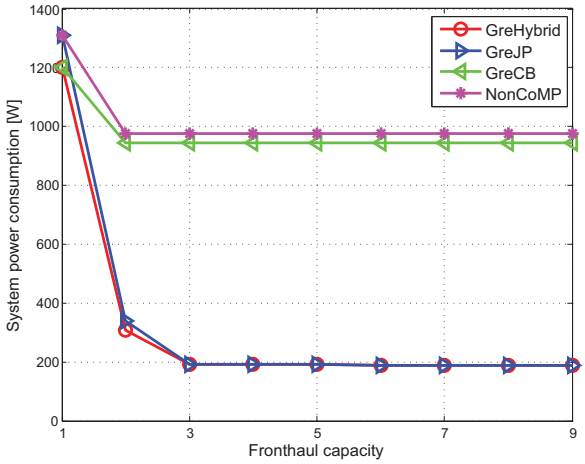

Fig. 8. System power consumption vs. S.

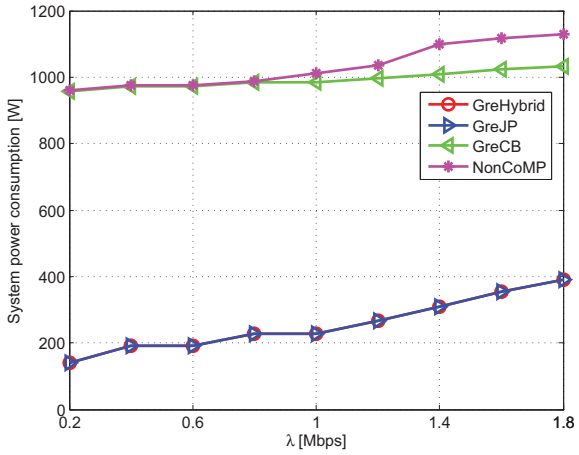

Fig. 10. System power consumption vs. λ.

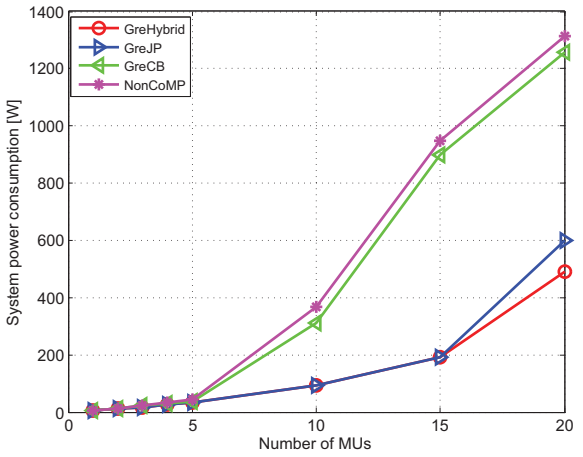

Fig. 9. System power consumption vs. number of MUs.

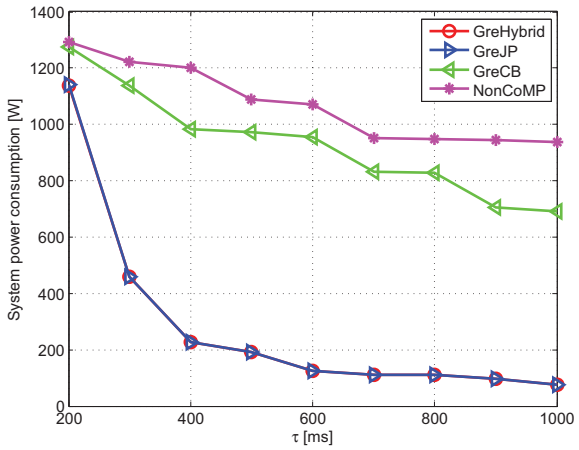

Fig. 11. System power consumption vs. τ.

is sufficient, the GreJP has better performance than the GreCB, for instance, $S \geq 2$. Note that the system power consumption is almost unchanged for all the schemes when $S \geq 3$. This is because the provisioned fronthaul capacity and DoF are excessive than the demand.

Next, Fig. 9 shows comparisons among the GreHybrid, GreJP, GreCB, and NonCoMP about system power consumption under different number of MUs. Here, $N = 8$, $S = 5$, $\lambda = 1$, and $\tau = 500$. It can be identified that the system power consumption increases with the number of MUs, and the GreHybrid and the NonCoMP consume the least and the most system power consumption respectively. Meanwhile, with given abundant fronthaul capacity, the GreJP outperforms the GreCB. Since less fronthaul capacity and DoF should be explored to support a small number of MUs ($K \leq 5$), the NonCoMP, GreCB, and GreJP can achieve the close system power consumption to the GreHybrid. Moreover, with increasing number of MUs, the GreHybrid can further explore fronthaul capacity and DoF to reduce processing power consumption at the cost of increasing transmit power consumption so as to reduce system power consumption significantly. However, to satisfy more MUs' QoS requirements, the other three schemes have to utilize increasing computational resources such that more system power is consumed.

In Figs. 10 and 11, we illustrate how λ and τ effect the system power consumption. Specifically, we set $N = 8$, $K = 15$, and $S = 5$. Besides, $\tau = 500$ and $\lambda = 1$ are adopted for Figs. 10 and 11, respectively. We can observe that with sufficient fronthaul capacity provisioned, the GreJP can approach the GreHybrid. Besides, the system power consumption increases with λ and decreases with τ. Note that for all the schemes, system power consumption with large λ and small τ are much higher than that with small λ and large τ. This is because plenty of computational resources are utilized under the strict conditions, i.e., large λ and small τ, resulting in much system power consumption.

VII. Conclusions

Taking the power consumption in the BBU pool and that at RRHs into consideration, this work has jointly optimized hybrid clustering and computation provisioning to minimize the system power consumption in downlink C-RAN subject to the limited fronthaul capacity. However, the system power minimization problem is integer non-linear and is difficult to solve. Therefore, we have exploited the special structure

of this problem, and have transformed it to an equivalent hybrid clustering problem, in which a series of VM assignment problems are embedded. Furthermore, we have optimally solved the hybrid clustering problem to achieve an optimal solution for system power minimization with high computational complexity and then, we have proposed a low complexity greedy algorithm to solve the hybrid clustering problem for practical implementation. Finally, extensive simulation results have exhibited performance improvements of our proposed algorithms against baseline algorithms in terms of system power consumption, which is one of the main merits of green C-RAN.

REFERENCES

[1] "C-RAN: The road towards green RAN," China Mobile, White Paper, Dec. 2013.

[2] T. Q. S. Quek, M. Peng, O. Simeone, and W. Yu, *Cloud Radio Access Networks: Principles, Technologies, and Applications*. Cambridge University Press, 2016.

[3] A. Checko, H. Christiansen, Y. Yan, L. Scolari, G. Kardaras, M. Berger, and L. Dittmann, "Cloud RAN for mobile networks - a technology overview," *IEEE Commun. Surveys & Tutorials*, vol. 17, no. 1, pp. 405–426, Mar. 2015.

[4] D. Gesbert, S. Hanly, H. Huang, S. Shamai Shitz, O. Simeone, and W. Yu, "Multi-cell MIMO cooperative networks: A new look at interference," *IEEE J. Sel. Areas Commun.*, vol. 28, no. 9, pp. 1380–1408, Dec. 2010.

[5] T. Werthmann, H. Grob-Lipski, and M. Proebster, "Multiplexing gains achieved in pools of baseband computation units in 4G cellular networks," in *Proc. IEEE PIMRC*, London, England, Sep. 2013, pp. 3328–3333.

[6] K. Guo, M. Sheng, J. Tang, T. Q. S. Quek, X. Wang, and Z. Qiu, "Cooperative transmission meets computation provisioning in downlink C-RAN," in *Proc. IEEE ICC*, Kuala Lumpur, Malaysia, May 2016, pp. 3709–3714.

[7] J. Tang, W. P. Tay, and T. Q. S. Quek, "Cross-layer resource allocation with elastic service scaling in cloud radio access network," *IEEE Trans. Wireless Commun.*, vol. 14, no. 9, pp. 5068–5081, Sep. 2015.

[8] V. N. Ha, L. B. Le, and N.-D. Dao, "Cooperative transmission in cloud RAN considering fronthaul capacity and cloud processing constraints," in *Proc. IEEE WCNC*, Istanbul, Turkey, Apr. 2014, pp. 1862–1867.

[9] M. Qian, W. Hardjawana, J. Shi, and B. Vucetic, "Baseband processing units virtualization for cloud radio access networks," *IEEE Wireless Commun. Letters*, vol. 4, no. 2, pp. 189–192, Apr. 2015.

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

[11] S. Luo, R. Zhang, and T. J. Lim, "Downlink and uplink energy minimization through user association and beamforming in C-RAN," *IEEE Trans. Wireless Commun.*, vol. 14, no. 1, pp. 494–508, Jan. 2015.

[12] J. Zhao, T. Q. S. Quek, and Z. Lei, "Coordinated multipoint transmission with limited backhaul data transfer," *IEEE Trans. Wireless Commun.*, vol. 12, no. 6, pp. 2762–2775, Jun. 2013.

[13] B. Dai and W. Yu, "Sparse beamforming and user-centric clustering for downlink cloud radio access network," *IEEE Access*, vol. 2, pp. 1326–1339, Oct. 2014.

[14] ——, "Energy efficiency of downlink transmission strategies for cloud radio access networks," *IEEE J. Sel. Areas Commun.*, vol. 34, no. 4, pp. 1037–1050, Apr. 2016.

[15] P. Patil and W. Yu, "Hybrid compression and message-sharing strategy for the downlink cloud radio-access network," in *Proc. IEEE ITA*, San Diego, CA, USA, Feb. 2014, pp. 1–6.

[16] S. H. Park, O. Simeone, O. Sahin, and S. Shamai, "Joint precoding and multivariate backhaul compression for the downlink of cloud radio access networks," *IEEE Trans. Signal Process.*, vol. 61, no. 22, pp. 5646–5658, Nov. 2013.

[17] T. X. Vu, H. D. Nguyen, and T. Q. S. Quek, "Adaptive compression and joint detection for fronthaul uplinks in cloud radio access networks," *IEEE Trans. Commun.*, vol. 63, no. 11, pp. 4565–4575, Nov. 2015.

[18] M. Hong, R. Sun, H. Baligh, and Z.-Q. Luo, "Joint base station clustering and beamformer design for partial coordinated transmission in heterogeneous networks," *IEEE J. Sel. Areas Commun.*, vol. 31, no. 2, pp. 226–240, Feb. 2013.

[19] C. Ng and H. Huang, "Linear precoding in cooperative MIMO cellular networks with limited coordination clusters," *IEEE J. Sel. Areas Commun.*, vol. 28, no. 9, pp. 1446–1454, Dec. 2010.

[20] J. Gong, S. Zhou, Z. Niu, L. Geng, and M. Zheng, "Joint scheduling and dynamic clustering in downlink cellular networks," in *Proc. IEEE GLOBECOM*, Houston, TX, USA, Dec. 2011, pp. 1–5.

[21] J. Zhang, R. Chen, J. Andrews, A. Ghosh, and R. Heath, "Networked MIMO with clustered linear precoding," *IEEE Trans. Wireless Commun.*, vol. 8, no. 4, pp. 1910–1921, Apr. 2009.

[22] J.-M. Moon and D.-H. Cho, "Inter-cluster interference management based on cell-clustering in network MIMO systems," in *Proc. IEEE VTC Spring*, Yokohama, Japan, May 2011, pp. 1–6.

[23] D. Liu, S. Han, C. Yang, and Q. Zhang, "Semi-dynamic user-specific clustering for downlink cloud radio access network," *IEEE Trans. Veh. Technol.*, vol. 65, no. 4, pp. 2063–2077, Apr. 2016.

[24] P. Rost, S. Talarico, and M. Valenti, "The complexity-rate tradeoff of centralized radio access networks," *IEEE Trans. Wireless Commun.*, vol. 14, no. 11, pp. 6164–6176, Nov. 2015.

[25] M. Valenti, S. Talarico, and P. Rost, "The role of computational outage in dense cloud-based centralized radio access networks," in *Proc. IEEE GLOBECOM*, Austin, TX, USA, Dec. 2014, pp. 1466–1472.

[26] L. Mashayekhy, M. Nejad, and D. Grosu, "Physical machine resource management in clouds: A mechanism design approach," *IEEE Trans. Cloud Comput.*, vol. 3, no. 3, pp. 247–260, Jul. 2015.

[27] A. Wiesel, Y. Eldar, and S. Shamai, "Zero-forcing precoding and generalized inverses," *IEEE Trans. Signal Process.*, vol. 56, no. 9, pp. 4409–4418, Sep. 2008.

[28] T. Yoo and A. Goldsmith, "On the optimality of multiantenna broadcast scheduling using zero-forcing beamforming," *IEEE J. Sel. Areas Commun.*, vol. 24, no. 3, pp. 528–541, Mar. 2006.

[29] R. Sun, H. Baligh, and Z.-Q. Luo, "Long-term transmit point association for coordinated multipoint transmission by stochastic optimization," in *Proc. IEEE SPAWC*, Darmstadt, Germany, Jun. 2013, pp. 330–334.

[30] Z. Shao, K. Guo, M. Sheng, S. Bian, Y. Zhang, J. He, Y. Li, and I. Chih-Lin, "Standards-compliant energy-saving schemes for downlink LTE/LTE-Advanced networks," in *Proc. IEEE PIMRC*, Washington, USA, Sep. 2014, pp. 86–90.

[31] N. Wang and T. Gulliver, "Queue-aware transmission scheduling for cooperative wireless communications," *IEEE Trans. Commun.*, vol. 63, no. 4, pp. 1149–1161, Apr. 2015.

[32] Y. Shi, J. Zhang, K. Letaief, B. Bai, and W. Chen, "Large-scale convex optimization for ultra-dense cloud-RAN," *IEEE Wireless Commun. Mag.*, vol. 22, no. 3, pp. 84–91, Jun. 2015.

[33] D. Bertsekas and R. Gallager, *Data Networks*, 2nd, Ed. New Jersey, U.S.: Prentice Hall, 1992.

[34] J. Liu, W. Chen, Z. Cao, and Y. J. Zhang, "Delay optimal scheduling for cognitive radios with cooperative beamforming: A structured matrix-geometric method," *IEEE Trans. Mobile Comput.*, vol. 11, no. 8, pp. 1412–1423, Aug. 2012.

[35] V. N. Ha and L. B. Le, "Joint coordinated beamforming and admission control for fronthaul constrained cloud-RANs," in *Proc. IEEE Globecom*, Austin, TX, USA, Dec. 2014, pp. 4054–4059.

[36] Y. Shi, J. Cheng, J. Zhang, B. Bai, W. Chen, and K. B. Letaief, "Smoothed l_p-minimization for green cloud-RAN with user admission control," *IEEE J. Sel. Areas Commun.*, vol. 34, no. 4, pp. 1022–1036, Apr. 2016.

[37] J. Zhao, T. Q. S. Quek, and Z. Lei, "Heterogeneous cellular networks using wireless backhaul: Fast admission control and large system analysis," *IEEE J. Sel. Areas Commun.*, vol. 33, no. 10, pp. 2128–2143, Oct. 2015.

[38] H. W. Kuhn, "The Hungarian method for the assignment problem," *Naval Research Logistics (NRL)*, vol. 2, pp. 83–97, Mar. 1955.

[39] L. Chen and N. Li, "On the interaction between load balancing and speed scaling," *IEEE J. Sel. Areas Commun.*, vol. 33, no. 12, pp. 2567–2578, Dec. 2015.

[40] Y. Wang, M. Sheng, X. Wang, L. Wang, W. Han, Y. Zhang, and Y. Shi, "Energy-optimal partial computation offloading using dynamic voltage scaling," in *Proc. IEEE ICC Workshop*, London,UK, Jun. 2015, pp. 2695–2700.

[41] A. P. Miettinen and J. K. Nurminen, "Energy efficiency of mobile clients in cloud computing," in *Proc. the 2nd USENIX Conference on Hot Topics in Cloud Computing (HotCloud)*, Boston, MA, USA, Jun. 2010, pp. 1–7.

[42] M.-H. Chen, B. Liang, and M. Dong, "A semidefinite relaxation approach to mobile cloud offloading with computing access point," in *Proc. IEEE SPAWC*, Stockholm, Sweden, Jun. 2015, pp. 186–190.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JSAC.2016.2624119, IEEE Journal on Selected Areas in Communications

Kun Guo received her B.E. degree in Telecommunications Engineering from Xidian University, Xi'an, China, in 2012, where she is currently working towards the Ph.D. degree in communication and information systems. She was a visiting student at the Singapore University of Technology and Design, Singapore, from Aug. 2015 to Jan. 2016. Her research interests focus on radio and computational resource management in cloud radio access network.

Min Sheng (M'03-SM'16) received the M.S. and Ph.D. degrees in Communication and Information Systems from Xidian University, Shaanxi, China, in 2000 and 2004, respectively. She is currently a Full Professor at the Broadband Wireless Communications Laboratory, the School of Telecommunication Engineering, Xidian University. Her general research interests include mobile ad hoc networks, wireless sensor networks, wireless mesh networks, third generation (3G)/4th generation (4G) mobile communication systems, dynamic radio resource management (RRM) for integrated services, cross-layer algorithm design and performance evaluation, cognitive radio and networks, cooperative communications, and medium access control (MAC) protocols. She has published 2 books and over 50 papers in refereed journals and conference proceedings. She was the New Century Excellent Talents in University by the Ministry of Education of China, and obtained the Young Teachers Award by the Fok Ying-Tong Education Foundation, China, in 2008.

Jianhua Tang (S'11-M'15) received his B.E. degree in Communication Engineering from Northeastern University, China, in 2010, and the Ph.D. degree in Electrical and Electronic Engineering from Nanyang Technological University, Singapore, in 2015. Currently, he is a Research Assistant Professor with the Department of Electrical and Computer Engineering at Seoul National University. He was a postdoctoral research fellow at the Singapore University of Technology and Design (SUTD) from Mar. 2015 to Oct. 2016. His research interests include cloud computing, content-centric network and cloud radio access network (C-RAN).

Tony Q.S. Quek (S'98-M'08-SM'12) received the B.E. and M.E. degrees in Electrical and Electronics Engineering from Tokyo Institute of Technology, Tokyo, Japan, respectively. At Massachusetts Institute of Technology, he earned the Ph.D. in Electrical Engineering and Computer Science. Currently, he is a tenured Associate Professor with the Singapore University of Technology and Design (SUTD). He also serves as the deputy director of the SUTD-ZJU IDEA. His main research interests are the application of mathematical, optimization, and statistical theories to communication, networking, signal processing, and resource allocation problems. Specific current research topics include heterogeneous networks, green communications, wireless security, internet-of-things, big data processing, and cognitive radio.

Dr. Quek has been actively involved in organizing and chairing sessions, and has served as a member of the Technical Program Committee as well as symposium chairs in a number of international conferences. He is serving as the Workshop Chair for IEEE Globecom in 2017 and the Special Session Chair for IEEE SPAWC in 2017. He is currently an Editor for the IEEE TRANSACTIONS ON COMMUNICATIONS and an Executive Editorial Committee Member for the IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS. He was Editor for the IEEE WIRELESS COMMUNICATIONS LETTERS, Guest Editor for the IEEE SIGNAL PROCESSING MAGAZINE (Special Issue on Signal Processing for the 5G Revolution) in 2014, and the IEEE WIRELESS COMMUNICATIONS MAGAZINE (Special Issue on Heterogeneous Cloud Radio Access Networks) in 2015. He is a co-author of the book "Small Cell Networks: Deployment, PHY Techniques, and Resource Allocation" published by Cambridge University Press in 2013 and the book "Cloud Radio Access Networks: Principles, Technologies, and Applications" by Cambridge University Press.

Dr. Quek was honored with the 2008 Philip Yeo Prize for Outstanding Achievement in Research, the IEEE Globecom 2010 Best Paper Award, the 2012 IEEE William R. Bennett Prize, the IEEE SPAWC 2013 Best Student Paper Award, the IEEE WCSP 2014 Best Paper Award, the IEEE PES General Meeting 2015 Best Paper, and the 2015 SUTD Outstanding Education Awards – Excellence in Research.

Zhiliang Qiu received the B.S. degree in communication engineering from the Northwestern Telecommunication Engineering Institute, Xi'an, China, in 1986, and the M.S. and Ph.D. degrees in the communication and information systems from Xidian University, Xi'an, China, in 1989 and 1999, respectively. He is currently a Full Professor with the School of Telecommunications Engineering, Xidian University. He is an expert in broadband information networks and Internet. His research interests focus on network framework, broadband access network, switching and avionics bus technologies.