LARGE-SCALE CONVEX OPTIMIZATION FOR ULTRA-DENSE CLOUD-RAN

YUANMING SHI, JUN ZHANG, KHALED B. LETAIEF, BO BAI, AND WEI CHEN

ABSTRACT

The heterogeneous cloud radio access network (Cloud-RAN) provides a revolutionary way to densify radio access networks. It enables centralized coordination and signal processing for efficient interference management and flexible network adaptation. Thus it can resolve the main challenges for next-generation wireless networks, including higher energy efficiency and spectral efficiency, higher cost efficiency, scalable connectivity, and low latency. In this article we will provide an algorithmic approach to the new design challenges for the dense heterogeneous Cloud-RAN based on convex optimization. As problem sizes scale up with the network size, we will demonstrate that it is critical to take unique structures of design problems and inherent characteristics of wireless channels into consideration, while convex optimization will serve as a powerful tool for such purposes. Network power minimization and channel state information acquisition will be used as two typical examples to demonstrate the effectiveness of convex optimization methods. Then we will present a two-stage framework to solve general large-scale convex optimization problems, which is amenable to parallel implementation in the cloud data center.

INTRODUCTION

With the dramatic increase in the number of smart mobile devices, and diversified wireless applications propelled by the advent of mobile social networks and the Internet of Things (IoT), we are in an era of a mobile data deluge. In particular, mobile data traffic has recently been doubling every year, which implies an astounding 1000 times increase in the following decade for 5G networks. Furthermore, new wireless applications bring new service requirements. For instance, intensive data services will be needed in crowded places such as stadiums and in densely populated metropolitan areas, while IoT applications call for scalable connectivity with diversified quality-of-service (QoS) requirements.

To meet these key requirements, a paradigm shift is needed in radio access networks. In particular, network densification supported by various types of techniques, including small cells [1], massive MIMO [2], and millimeter-wave communications [3], has been proposed as a promising approach, which will result in a dense heterogeneous network. With improved spatial reuse and traffic dependent deployment, dense networks have the potential to boost network capacity and provide diversified wireless services. However, there are formidable challenges to achieving the benefit of dense heterogeneous networks, including the magnified interference issue, the high capital expenditure (CAPEX) and operating expenditure (OPEX), mobility management, etc. Therefore, a holistic approach to deploy and manage dense networks is required, and efficient multi-tier collaboration should be supported.

The heterogeneous cloud radio access network (Cloud-RAN) [4] is a promising centralized radio access technology to address the key challenges toward network densification by leveraging recent advances in cloud-computing technology. In particular, intra-tier and inter-tier interference can be effectively mitigated by centralized signal processing and coordination at the cloud data center. Furthermore, with elastic network reconfiguration and adaptation, the operation efficiency of the heterogeneous Cloud-RAN can be significantly improved. For example, by adaptively switching on/off radio access points, and adjusting computing resources at the cloud data center, the network can be well adapted to spatial and temporal traffic fluctuations.

However, the dense heterogeneous Cloud-RAN brings new design challenges, mainly due to the enlarged problem size as the design parameters and the required side information grow substantially. In this article we will provide a holistic viewpoint for designing dense Cloud-RAN via convex optimization. It has been well recognized that convex optimization provides an indispensable set of tools for designing wireless communication systems [5], e.g. coordinated beamforming, power control, user admission control, as well as data routing and flow control. The main reason for the success of convex optimization lies in its capability of flexible formulations, efficient globally optimal algorithms, e.g. the interior-point method, and the ability to leverage convex analysis to explore the solution structure, e.g. the uplink-downlink duality in the

Yuanming Shi, Jun Zhang and Khaled B. Letaief are with The Hong Kong University of Science and Technology.

Bo Bai and Wei Chen are with Tsinghua University.

1536-1284/15/$25.00 © 2015 IEEE

multiuser beamforming problem. However, in dense Cloud-RAN, with its complex architecture, as well as the large size of optimization variables and parameters, new challenges arise.

In this paper we will present new convex optimization methods for dense Cloud-RAN, considering three key aspects in such networks. First, a convex relaxation approach will be shown to be a powerful tool to deal with design problems with complicated variables, including both discrete and continuous variables. Such problems arise frequently in dense collaborative networks, such as the network power minimization problem. We will then consider channel state information (CSI) acquisition in dense Cloud-RAN, which is critical for centralized signal processing and resource allocation. A convex regularized optimization approach is used to exploit the channel structure for the high-dimensional channel estimation problem, while a successive convex approximation algorithm is used for designing stochastic coordinated beamforming to deal with CSI uncertainty. Major design problems in dense Cloud-RAN all involve a large number of parameters and design variables, and our third consideration is on large-scale convex optimization algorithms. A general two-stage approach is proposed, which can scale to large problem sizes and exploit the parallel computing environment in the cloud data center.

Network Architecture and Research Challenges

In this section we will introduce the main entities of heterogeneous Cloud-RAN, including the cloud data center, the mobile hauling network, and the radio access network, followed by an overview of new research challenges.

Network Architecture

Heterogeneous Cloud-RAN is a disruptive technology that takes advantage of recent advances in cloud-computing to revolutionize next-generation wireless networks. Its architecture is shown in Fig. 1. A key feature of heterogeneous Cloud-RAN is that different radio access points will be equipped with different entities, including communication, computation, and storage units. Thus we will also call it a *multi-entity next generation radio access network*, or MENG-RAN. Specifically, the cloud data center serves as a central cloud infrastructure for the dense heterogeneous radio access network consisting of different types of access points. The key advantage of the heterogeneous Cloud-RAN lies in the centralized coordination at the cloud data center, supported by the mobile hauling network to transfer the information to and from different access points. In the following, we will introduce the main functionality of each entity, as well as some deployment issues.

Cloud Data Center: The cloud data center consists of shared and reconfigurable computation and storage resources. Thanks to such a shared hardware platform, both the CAPEX (e.g. via low-cost site construction) and OPEX (e.g. via centralized cooling), as well as the management effort, can be significantly reduced. Besides performing basic baseband digital signal processing for transmission and reception, the cloud data center can also provide cloud-computing functionalities, such as on-demand services via virtualization with multiple virtual machines and resource pooling, and parallel computing for scalable algorithm implementation.

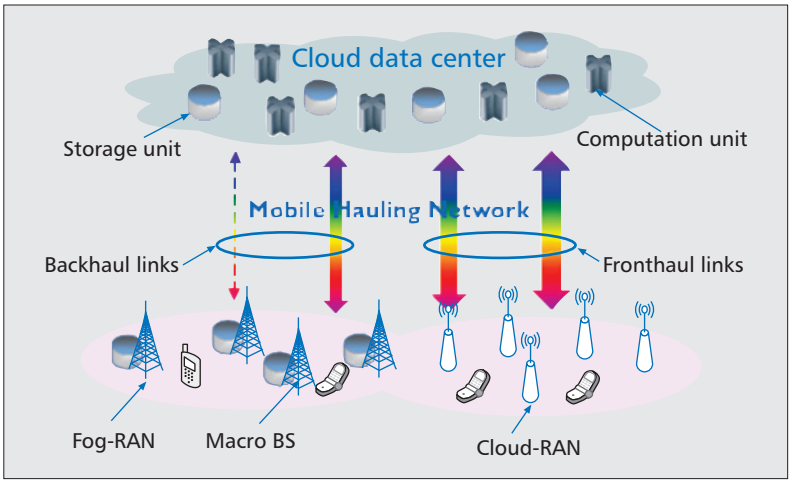

Figure 1. The architecture of heterogeneous Cloud-RAN, also called as MENG-RAN. It consists of different types of radio access networks, including Cloud-RAN and Fog-RAN.

The centralized signal processing enabled by the cloud data center is essential for the performance gains of Cloud-RAN. Specifically, with densely deployed remote radio heads (RRHs), by applying advanced signal processing algorithms in the computationally powerful cloud data center, large-scale cooperation can be achieved, thereby improving both spectral efficiency and energy efficiency. Moreover, with centralized coordination, effective dynamic resource allocation can be provided to smooth out spatial and temporal traffic fluctuations.

Radio Access Network: In heterogeneous Cloud-RAN, access points can be divided into two categories, i.e. low-cost low-power RRHs and powerful macro BSs. Specifically, each RRH only consists of a passband signal processor, an amplifier, and an A/D converter to support basic transmission and reception functionality, while the baseband signal processing is carried in the cloud data center. The data transmitted between the cloud data center and RRHs are typically oversampled real-time I/Q digitalized baseband streams in the order of Gb/s [6]. For such access nodes, the mobile hauling network that provides high-capacity connection to the cloud data center is usually called the mobile fronthaul network. The typical requirements of fronthaul links are: link capacity (1–10 Gb/s), latency (≤ 0.1 ms), and distance (1–10 km).

On the other hand, the compact macro BSs are with additional baseband signal processors and storage units. In this way, the computation and storage resources are pushed to the edge of the radio access network, which we will call a Fog-RAN. In this scenario the data transmitted between the cloud data center and BSs using the packet-based interface are only side information

A key challenge of heterogeneous Cloud-RAN is to transfer the data traffic between the cloud data center and the radio access network by the mobile hauling network. The capacity of each mobile hauling link will affect both the network performance and deployment cost, so it should be carefully picked.

and user messages in the order of several hundred Mb/s. The mobile hauling network for such nodes is called the mobile backhaul network, which has a low capacity requirement. The typical requirements of backhaul links are: link capacity (200–500 Mb/s), latency (≤ 10 ms), and distance (≤ 1 km). By caching the content in the storage units of BSs during the idle hours, the backhaul signaling overhead can be further reduced. However, with limited computational capability at each BS, the cooperative gain in such a scenario will be lower than RRHs.

Mobile Hauling Network: A key challenge of heterogeneous Cloud-RAN is to transfer the data traffic between the cloud data center and the radio access network by the mobile hauling network. The capacity of each mobile hauling link will affect both the network performance and deployment cost, so it should be carefully picked.

Specifically, for the mobile fronthaul network connecting RRHs and the cloud data center, there is a stringent requirement on latency and synchronization, as well as low jitter and error tolerance [6]. Both the low-cost wavelength-division multiplexing passive optical network (WDM-PON) and orthogonal frequency-division multiple access passive optical network (OFDMA-PON) are promising candidates. With much lower capacity and latency requirements of mobile backhaul networks, the mmW technology (V-band frequencies, i.e. 40–75 GHz, and E-band frequencies, i.e. 71–76 GHz and 81–86 GHz) [3] serves as a cost-effective backhaul solution.

Research Challenges

The new architecture of MENG-RAN will bring new opportunities as well as new design challenges. In this paper we will address the main design challenges of dense Cloud-RAN from a unique perspective, i.e. we will focus on convex optimization based approaches. Convex optimization has long been recognized as a powerful tool for designing wireless networks [5]. In dense Cloud-RAN, as the design problems are entering a new regime with high-dimensional optimization variables and a large number of parameters, new design challenges arise. We will demonstrate the strength of convex optimization by developing new methodologies for the key design problems. In particular, the main focus will be on the following areas.

- In dense Cloud-RAN with a large number of RRHs, it is critical to select RRHs to adapt to the temporal and spatial data dynamics, thereby improving the operating efficiency. To enable such adaptation, the new challenge comes from the composite design variables, which consist of both discrete and continuous variables for RRH selection and beamforming design, respectively. This often yields a mixed integer nonlinear programming problem and is NP-hard. With network power minimization as a representative example, we will introduce the group sparse beamforming algorithm [7] to efficiently solve such problems.
- CSI is essential to various cooperation strategies of dense Cloud-RAN, but its acquisition becomes challenging as a large number of access points are involved in cooperation. To address the channel estimation challenge with limited training resources, a convex regularized optimization approach [8] will be discussed in a later section to exploit unique structures of wireless channels. To further deal with the resulting CSI uncertainty, a successive convex approximation algorithm is proposed to solve the corresponding highly intractable stochastic coordinated beamforming problem [9].
- In dense Cloud-RAN, the cloud data center will typically support hundreds of RRHs [6], and thus all the optimization algorithms need to scale to large problem sizes. Furthermore, for many design problems, optimization algorithms should have the capability to detect infeasibility accurately. To meet both requirements, we will present a two-stage large-scale convex optimization framework [10] to leverage the cloud-computing environment in the cloud data center. In particular, the heterogeneous Cloud-RAN will further present new challenges on the synchronization and distributed implementation of large-scale optimization algorithms.

Convex Optimization Methods for Cloud-RAN

In this section we will present three design methods based on convex optimization to address the new research challenges in dense Cloud-RAN. In particular, we will demonstrate that convex optimization has the advantage of enabling flexible formulations and scalable algorithms, which will make it a powerful design tool for MENG-RAN. The main idea of the three proposed methods is illustrated in Table I. While for illustrative purposes we mainly focus on the Cloud-RAN, the proposed methods are generic and applicable to different coordination strategies (e.g. with CSI sharing only) among RRHs and macro BSs in heterogeneous Cloud-RAN.

Group Sparse Beamforming: Convex Relaxation for Network Power Minimization

With densely deployed access points in heterogeneous Cloud-RAN, effective network adaptation is critical to improve the resource utilization efficiency. In particular, considering the spatial and temporal traffic fluctuation, we can dynamically select appropriate access points and mobile hauling links to serve active users. In this way, we can reduce the power consumption of both the access points and hauling links, and also reduce the signaling overhead. However, such adaptive operation will present unique challenges for network optimization, and the design problem will need to handle both discrete (e.g. for RRH selection and user data routing) and continuous (e.g. beamforming coefficients) variables. In this section we will use the network power minimization problem as an example to illustrate a powerful method based on convex relaxation to deal with such design problems.

In Cloud-RAN, the network power consumption consists of two main components: the transmit power of the active RRHs, and the power consumption of their fronthaul links. By adap-

	Group sparse beamforming with convex relaxation	Convex optimization for CSI estimation and exploitation	Large-scale convex optimization
Problem statement	*Network Power Minimization Problem:* min Fronthaul link power +RRH transmit power s.t. Per mobile user QoS constraints Per RRH power constraints.	*High-Dimensional Channel Extimation Problem:* min Loss function + regularizing function *Stochastic Coordinatd Beamforming Problem:* min Total transmit power s.t. Probabilistic QoS constraints Per RRH power constraints.	*Standard Cone Program:* min $\mathbf{c}^T\mathbf{v}$ s.t. $\mathbf{A}\mathbf{v} + \boldsymbol{\mu} = \mathbf{b}$ $(\mathbf{v}, \boldsymbol{\mu}) \in \mathbb{R}^n \times \mathcal{K}$. Problem data: **A**, **b**, **c**.
Algorithm description	*Workflow:* **Step one:** group sparse including norm minimization. **Step two:** fronthaul link and RRH selection. **Step three:** transmit beamforming design.	*Workflow:* **Step one:** CSI selection: determine "relevant" channel links. **Step two:** downlink training and/or uplink feedback. **Step three:** stochastic coordinated beamforming design with mixed CSI.	*Workflow:* **Stage one:** standard cone program transformation. **Stage two:** the standard form problem is solved by an ADMM-based convex solver.
	Explanations: 1) In step one, the (approximated) group sparse beamforming vector is obtained by solving a convex group sparse inducing norm minimization problem. 2) Optimal active RRHs can be obtained after step two by solving a sequence of feasibility problems to check if the remaining active RRHs can support the QoS rquirements. 3) In step three, optimal coordinated beamforming coefficients are obtained for the active RRHs by solving a convex program.	*Explanations:* 1) A practical way to determine the optimal indices of the "relevant" channel links is to exploit the sparsity of large-scale fading coefficients [9]. 2) In step two, a convex regularized optimization method is adopted to reduce the training overhead by exploiting the channel spatial and temporal prior information. 3) Only the distribution information in the mixed CSI is required to solve the stochastic coordinated beamforming problem in step three.	*Explanations:* 1) Most general convex programs can be transformed into the standard cone program form with κ as the Cartesian product of the standard cones, e.g., second-order cone. 2) An operator splitting method based on ADMM algorithm is adopted to solve the standard cone programs.

Table 1. Convex optimization methods for dense cloud-RAN.

tively switching off some RRHs, we can save the power consumption of the corresponding fronthaul links, and thus reduce the network power. Subsequently, the network power consumption will become a composite function, i.e. the transmit power component is determined by the continuous beamforming vectors, while the fronthaul power component depends on the active RRH set, which is discrete. Thus the network power minimization problem becomes a mixed integer nonlinear programming (MINLP) problem, which is NP-hard. There are lots of other design problems in MENG-RAN that have similar structures, for which a powerful tool based on convex relaxation has recently been proposed. In [7], a three-step group sparse beamforming (GSBF) algorithm is proposed to minimize the network power consumption by adaptively selecting RRHs via controlling the group-sparsity structure of the aggregative beamforming vector, as illustrated in Table 1. In particular, the simulation results in Fig. 2 demonstrate that the proposed group sparse beamforming algorithm achieves near-optimal performance. It also shows the importance of adaptively selecting RRHs to improve the energy efficiency in dense Cloud-RAN.

The main idea of group sparse beamforming is to convexify the non-convex composite objective function via convex relaxation. In particular, the weighted l_1/l_2-norm of the aggregative beamforming vector of all the RRHs is shown to be the tightest convex lower bound of the original objective function. It is then used to induce the group sparsity pattern of the beamforming vector, thereby providing guidelines for RRH selection. This approach essentially exploits the group sparsity structure of the optimal solution. Specifically, all the beamforming coefficients of one RRH can be regarded as a group. When an RRH is switched off, the corresponding beamforming coefficients in the same group will be set to be zeros simultaneously. Overall, as there will be multiple RRHs being switched off to save the power, the optimal aggregative beamforming vector will thus have a group sparsity structure.

For other network performance optimizations in MENG-RAN that need to jointly allocate combinatorial resources (e.g. RRH selection, data assignment, user scheduling and subcarrier allocation) and optimize continuous resources (e.g. beamforming coefficients, power allocation), group sparse beamforming provides a principled way to develop polynomial-time complexity algorithms. In particular, such a convex relaxation approach helps leverage the problem structures (e.g. group sparsity) via carefully deriving corresponding convex surrogates (e.g. l_1/l_2-norm) for the original non-convex functions (e.g. the composite objective function). Efficient convex optimization algorithms can then be applied. Encouraging progress has been made in applying this approach to wireless networks, e.g. for uplink and downlink energy minimization in Cloud-RAN [11] and for data assignment in backhaul links [12]. Meanwhile, there remain a variety of interesting open questions:

- Sparsity inducing norms for more general settings and more general utility functions should be derived, e.g. considering imperfect CSI, multicast transmission, limited fronthaul link capacity, and energy efficiency, which is defined as the ratio between the achievable sum rate and the total network power consumption.
- Fast infeasibility detection algorithms need to be developed to speed up the selection procedure (e.g. RRH selection), as a sequence of feasibility problems are needed to be solved to make a final decision.

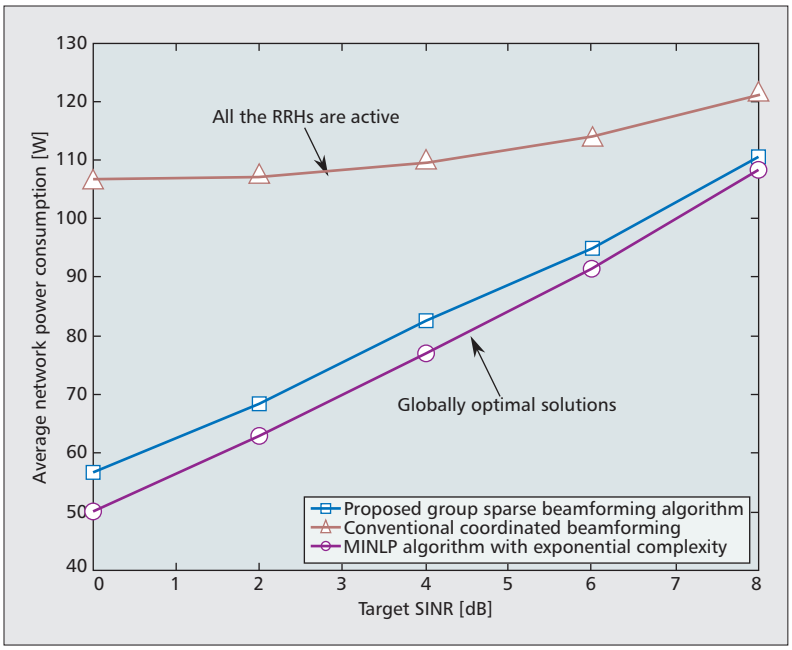

Figure 2. Average network power consumption versus target signal-to-interference-plus-noise ratio (SINR). Simulation details: 10 2-antenna RRHs and 15 single-antenna mobile users uniformly and independently distributed in the square region [–1000, 1000] × [–1000, 1000] meters, the relative fronthaul link power consumption (i.e., the power saving by switching off one fronthaul link and the corresponding RRH) is set to be $(5 + l)$W, $l = 1, \ldots, 10$. Standard cellular channel model is adopted [7].

- The optimality of the group sparse beamforming algorithm should be characterized, which will be of a significant theoretical value.

Convex Optimization for CSI Acquisition and Exploitation

CSI plays a pivotal role for effective interference management and resource allocation as well as enabling scalable and flexible cooperation in MENG-RAN, e.g. for network adaptation. With dense deployment of access points, CSI acquisition becomes a formidable task. In particular, due to the limited radio resources for CSI training, the training pilot length is typically smaller than the dimension of the channel. Conventional methods, such as the least square estimate, become inapplicable in such settings, and novel CSI acquisition methodologies are needed. A unique property of Cloud-RAN with geographically distributed RRHs is the sparsity of the large-scale fading coefficients due to path loss. That is, the channel links between the RRHs and MUs that are far away will have negligible channel gains and contribute little to system performance. A practical way to reduce the CSI acquisition overhead in terms of training and/or feedback is to only obtain a subset of the strongest channel links. This is called compressive CSI acquisition [9], in which only a subset of "relevant" channel links (e.g. the instantaneous channel links with the dominated large-scale fading coefficients) will be obtained. Furthermore, it is critical to adjust the beamforming design so that it can effectively exploit the available channel information, while taking its uncertainty into consideration.

To address the above challenges, in this section we will provide a unified convex optimization based framework to develop a high-dimensional channel estimation algorithm and a successive convex approximation algorithm to deal with the CSI uncertainty, as shown in Table 1.

Convex Regularized Optimization for High-Dimensional Channel Estimation: It is well known that exploiting the "low-complexity" structures of channels can reduce the training overhead, e.g. by exploiting channel sparsity via compressed sensing. However, in dense Cloud-RAN with a large number of access points and limited radio resources, the conventional compressed sensing based approach may lose its effectiveness as it only exploits channel sparsity as a prior. This motivates us to further exploit the temporal correlation of the channels across different fading blocks to enhance the estimation performance.

Specifically, a convex regularized optimization approach [8] can be adopted to convert physical notions of channel prior information or structures (i.e. heterogenous large-scale fading and temporal correlation) into appropriate convex regularizing functions, thereby solving the underdetermined channel estimation problem with insufficient training pilots. Thus the channel estimation problem is formulated as a convex optimization problem, for which the objective consists of two parts: a loss function to measure the compatibility between the estimate and the observation, and a convex regularizing function that encodes the prior information of the channel structure. Figure 3a demonstrates the channel estimation performance with different available prior information. It shows that the training overhead can be reduced by exploiting the spatial and temporal prior information. This regularized approach has the potential to incorporate other structures of wireless channels for training overhead reduction. Furthermore, such convex optimization formulation also enables efficient and scalable algorithm design, as will be shown later.

Successive Convex Approximation for Stochastic Coordinated Beamforming: With compressive CSI acquisition, the obtained CSI is actually of mixed types, consisting of a subset of imperfect instantaneous CSI and statistical CSI for the other channel links that have not been trained. To exploit the available mixed CSI, a new beamforming approach is needed to alleviate the performance degradation due to CSI uncertainty. In [9] a stochastic coordinated beamforming framework based on joint chance constrained programming is proposed to deal with CSI uncertainty. A probabilistic QoS constraint is adopted, i.e. outage is allowed for each user, but its probability should be below a given threshold. Such a probabilistic constraint is motivated by the fact that most wireless networks can tolerate occasional outages.

Although it can consider different types of CSI uncertainties, the stochastic coordinated beamforming problem is highly intractable, and thus only suboptimal algorithms exist. For instance, the scenario approach intends to approximate the probabilistic QoS constraint by multiple "sampling" constraints using the Monte

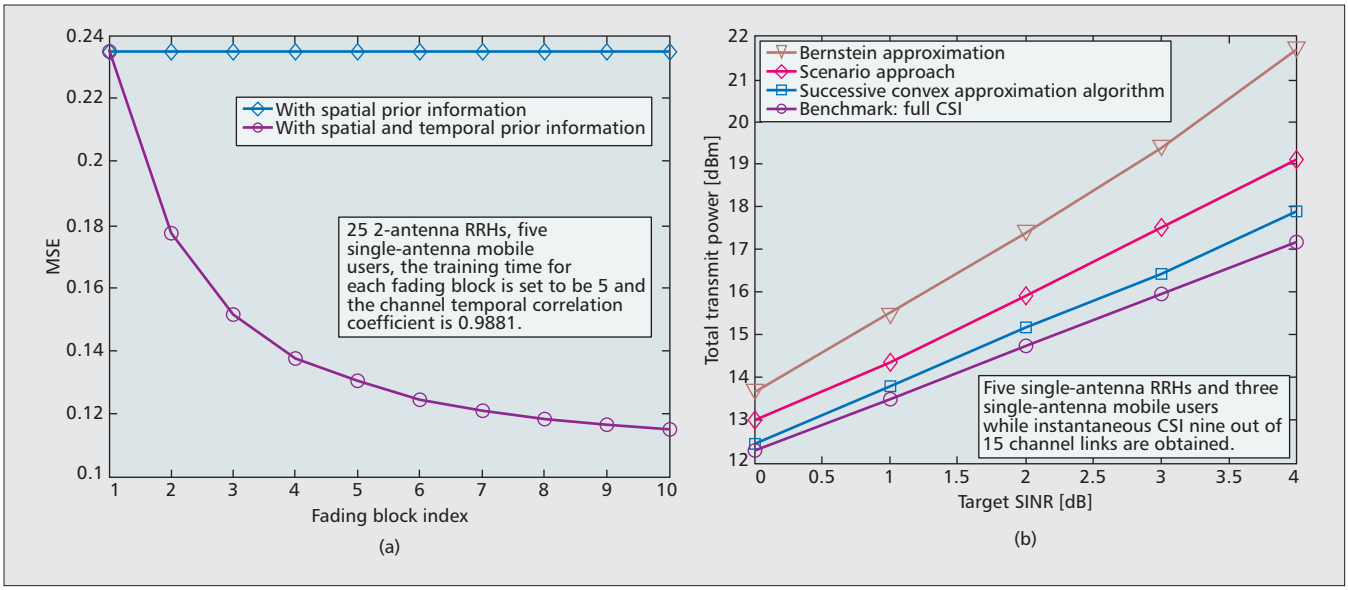

Figure 3. Convex optimization approach for CSI estimation and exploitation. In (a), we simulate the convex regularized optimization for high-dimensional channel estimation with least-squares as the loss function and weighted l_1-norm (where the weights are defined as the inverse of the corresponding large-scale fading coefficients) and squared l_2-norm as the regularizing functions to exploit the spatial and temporal prior information, respectively. The performance metric is given by the mean squared error $\text{MSE} = \mathbb{E}[||\hat{\mathbf{H}} - \mathbf{H}||_F^2]$. We assume that the channel follows the first-order stationary Gauss-Markov process and we consider such a random process of length 10 blocks for each channel estimation procedure. In particular, the channel in the i-th block is estimated based on all the received training signals at blocks $k = 1, \ldots, i$. In (b), we simulate the stochastic coordinated beamforming problem with mixed CSI to minimize the total transmit power while satisfying the probabilistic QoS constraints and per RRH power constraints.

Carlo simulation, while the Bernstein approximation method tries to bound the chance constraint with a closed-form (please refer to [9, Section IV] for more details), but both algorithms often yield suboptimal solutions, as shown in Fig. 3b. To avoid performance loss due to suboptimal solutions, a novel successive convex approximation algorithm is proposed in [9] with an optimality guarantee. It can also help us investigate the effectiveness of the proposed CSI acquisition methods. As shown in Fig. 3b, the proposed partial CSI acquisition method with stochastic coordinated beamforming can provide good performance while significantly reducing CSI overhead. Furthermore, this figure shows that there is a tradeoff between the performance and the computational complexity, e.g. the Bernstein approximation method has the lowest computational complexity but it yields the highest transmit power.

From the above discussion, we see that exploiting the low-complexity channel structure becomes the "blessing" to overcome the "curse of dimensionality" for massive CSI acquisition in dense Cloud-RAN. In particular, convex regularized optimization provides a flexible and computationally efficient way to exploit the channel structures to reduce the training overhead. Stochastic beamforming with the convex approximation algorithm, on the other hand, provides a flexible and optimal way to deal with CSI uncertainty. Note that the proposed CSI acquisition and exploitation method is generic and can be applied for different coordination strategies in heterogeneous Cloud-RAN, as it only depends on the channel properties (e.g. sparsity) for CSI acquisition and available CSI (e.g. partial and imperfect CSI) for precoding and decoding strategy design. There are, however, several interesting problems that need to be further investigated:

- The statistical performance of the convex regularized channel estimation method should be carefully evaluated.
- Scalable algorithms (e.g. the stochastic ADMM algorithm) need to be developed to solve the corresponding stochastic optimization problems to deal with CSI uncertainty.
- The fundamental performance limits of such dense distributed wireless cooperative networks with specific CSI acquisition methods and precoding/decoding strategies should be investigated.

Parallelizable Large-Scale Convex Optimization Algorithms

We have demonstrated that convex optimization based approaches are powerful for various design problems in dense Cloud-RAN, where a convex problem (e.g. the convex regularized optimization based channel estimation problem) or a sequence of convex subproblems (e.g. the three-stage GSBF algorithm) need to be solved. However, the drastically increased network density places tremendous pressure on the computational efficiency of the algorithms. For instance, for a Cloud-RAN with 100 single-antenna RRHs and 100 single-antenna MUs, the dimension of the aggregative coordinated beamforming vector (i.e. the optimization variables) will be 10^4, while solving convex quadratic programs has cubic complexity using the interior-point method.

Open issues of particular interests include the optimization of the virtual machine placement and resource utilization in the virtualized cloud data center, multiterminal baseband signal compression with arbitrary topology fronthaul networks, and data flow in backhaul networks.

Network size ($L = K$)		20	50	100	150	200
CVX + SDPT3	Modeling time [sec]	0.7563	4.4301	N/A	N/A	N/A
	Solving time [sec]	4.2835	326.2513	N/A	N/A	N/A
	Objective [W]	12.2488	6.5216	N/A	N/A	N/A
Matrix stuffing + SCS	Modeling time [sec]	0.0128	0.2401	2.4154	9.4167	29.5813
	Solving time [sec]	0.1009	2.4821	23.8088	81.0023	298.6224
	Objective [W]	12.2523	6.5193	3.1296	2.0689	1.5403

Table 2. Time and solution results with different convext optimization frameworks for the coordinated beamforming problem to minimize the total transmit power with L 2-antenna RRHs and K single-antenna MUs.

Moreover, a sequence of convex feasibility problems need to be solved for lots of design problems, e.g. for the RRH selection in the network power minimization problems. However, most existing custom algorithms, e.g. the ADMM based algorithms [13] and the uplink-downlink duality approach, cannot provide the certificates of infeasibility. Thus, effective feasibility detection should be embedded in the algorithm. To resolve these challenges, we provide a generic two-stage approach in [10], as illustrated in Table 1. It can solve large-scale general convex optimization problems in parallel, with the capability of returning either the optimal solution or the certificate of infeasibility. The time complexity of solving a general optimization problem involves two parts, i.e. the modeling time that transforms the problem to a standard form, and the solving time that solves the standard form problem. We will next demonstrate that the proposed approach can help to significantly improve the efficiency of both parts, and thus it can scale to large problem sizes.

In the first stage, the original convex program is transformed into a standard cone program with only a subspace constraint and a convex cone constraint formed by the Cartesian product of symmetric cones (e.g. second-order cones and positive semidefinite cones). This procedure can be done very fast using the matrix stuffing technique, which only needs to copy the problem data into the pre-stored structure of the standard cone program. As a result, all the structure information of the original convex problem is packed into the constraints. In the second stage, the structure of the standard form problem will be exploited to enable parallel computing and infeasibility detection using the alternative direction method of multipliers (ADMM) algorithm [14]. Such a first-order method features low or medium accuracy solutions, and nearly dimension-independent convergence rates. While the first stage will significantly reduce the modeling time, the second stage will help reduce the solving time.

Simulation results in Table 2 demonstrate the speedups of several orders of magnitude over the state-of-the-art modeling framework CVX and interior-point solvers (e.g. SDPT3). Specifically, for the modeling time, this table shows that the proposed matrix stuffing technique can speed up about 20 to 60 times compared to the parser/solver modeling framework CVX. For the solving time, taking $L = 50$ as an example, the ADMM based solver SCS can speedup 130 times over the inter-point solver SDPT3, which is inapplicable for large-scale problems with $L \geq 100$. Therefore, the proposed generic two-stage based large-scale convex optimization framework scales well to large-scale performance optimization problems based on the convex optimization approach in heterogeneous Cloud-RAN. This enables us to investigate the performance of optimal beamforming strategies in large-scale networks, which demonstrate significant gains over suboptimal transmission strategies, as shown in Fig. 4. This clearly shows the importance of developing optimal beamforming algorithms for such dense cooperative networks.

To better exploit the advantage of the proposed approach, further efforts are needed, specified as follows.

- Communication constraints and synchronization among hardware computation units should be considered to practically implement parallel algorithms.
- Efficient and cheap subspace and cone projection (especially for the semidefinite cone projection) algorithms are needed in the ADMM algorithm for solving large-scale standard cone programs.
- Other first-order methods (e.g. the subgradient method and the Frank-Wolfe (projection free) method) are also worth considering to exploit the problem structures. This, in contrast, may be achieved by packing all the structure information into the objective function, leaving only linear constraints.
- To enable distributed implementation of the ADMM algorithm among macro BSs in heterogeneous Cloud-RAN, it is critical to exploit structures (e.g. decomposability [15]) of the resulting problems for different coordination strategies.

Conclusions and Discussions

This article discussed the benefits and design challenges of the dense heterogeneous Cloud-RAN (also called MENG-RAN). Convex optimization methods were demonstrated to be

powerful tools to address the key design challenges by effectively exploiting problem structures and network properties. The presented results stand for a new paradigm for designing dense Cloud-RAN, and further investigation will be needed. Other open issues of particular interests include the optimization of the virtual machine placement and resource utilization in the virtualized cloud data center, multiterminal baseband signal compression with arbitrary topology fronthaul networks, and data flow in backhaul networks.

Acknowledgments

This paper is partially supported by the Hong Kong Research Grant Council under Grant No. 16200214, the National Basic Research Program of China (973 Program) No. 2013CB336600, and the NSFC Excellent Young Investigator Award No. 61322111.

References

[1] T. Q. Quek *et al.*, *Small Cell Networks: Deployment, PHY Techniques, and Resource Management*. Cambridge University Press, 2013.
[2] F. Rusek *et al.*, "Scaling up MIMO: Opportunities and Challenges with Very Large Arrays," *IEEE Signal Proc. Mag.*, vol. 30, no. 1, 2013, pp. 40–60.
[3] S. Rangan, T. Rappaport, and E. Erkip, "Millimeter-Wave Cellular Wireless Networks: Potentials and Challenges," *Proc. IEEE*, vol. 102, pp. 366–85, Mar. 2014.
[4] M. Peng *et al.*, "Heterogeneous Cloud Radio Access Networks: A New Perspective for Enhancing Spectral and Energy Efficiencies," *IEEE Wireless Commun. Mag.*, vol. 21, Dec. 2014, pp. 126–35.
[5] D. P. Palomar and Y. C. Eldar, *Convex Optimization in Signal Processing and Communications*, Cambridge University Press, 2010.
[6] China Mobile, "C-RAN: The Road Towards Green RAN," White Paper, ver. 3.0, Dec. 2013.
[7] Y. Shi, J. Zhang, and K. B. Letaief, "Group Sparse Beamforming for Green Cloud-RAN," *IEEE Trans. Wireless Commun.*, vol. 13, May 2014, pp. 2809–23.
[8] M. J. Wainwright, "Structured Regularizers for High-Dimensional Problems: Statistical and Computational Issues," *Annu. Rev. Stat. Appl.*, vol. 1, 2014, pp. 233–53.
[9] Y. Shi, J. Zhang, and K. Letaief, "Optimal Stochastic Coordinated Beamforming for Wireless Cooperative Networks with CSI Uncertainty," *IEEE Trans. Signal Proc.*, vol. 63, Feb. 2015, pp. 960–73.
[10] Y. Shi, J. Zhang, and K. B. Letaief, "Scalable Coordinated Beamforming for Dense Wireless Cooperative Networks," *Proc. IEEE Global Communications Conf. (GLOBECOM)*, Austin, TX, 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, Jan. 2015, pp. 494–508.
[12] B. Dai and W. Yu, "Sparse Beamforming and User-Centric Clustering for Downlink Cloud Radio Access Network," *IEEE Access*, vol. 2, Nov. 2014, pp. 1326–39.
[13] W.-C. Liao *et al.*, "Base Station Activation and Linear Transceiver Design for Optimal Resource Management in Heterogeneous Networks," *IEEE Trans. Signal Proc.*, vol. 62, Aug. 2014, pp. 3939–52.
[14] S. Boyd *et al.*, "Distributed Optimization and Statistical Learning via the Alternating Direction Method of Multipliers," *Found. Trends in Mach. Learn.*, vol. 3, July 2011, pp. 1–122.
[15] M. Chiang *et al.*, "Layering as Optimization Decomposition: A Mathematical Theory of Network Architectures," *Proc. IEEE*, vol. 95, no. 1, 2007, pp. 255–312.

Biographies

YUANMING SHI [S'13] (yshiac@ust.hk) received his B.S. degree in electronic engineering from Tsinghua University, Beijing, China, in 2011. He is currently working toward the Ph.D. degree in the Department of Electronic and Computer Engineering at the Hong Kong University of Science and Technology (HKUST). His research interests include large-scale optimization and analysis, stochastic modeling and optimization, Riemannian optimization, high-dimensional statistics, and wireless networking.

JUN ZHANG [M'10] (eejzhang@ust.hk) received the Ph.D. degree in electrical and computer engineering from the University of Texas at Austin in 2009. He is currently a research assistant professor in the Department of Electronic and Computer Engineering at the Hong Kong University of Science and Technology. Dr. Zhang co-authored the book *Fundamentals of LTE* (Prentice-Hall, 2010). His research interests include wireless communications and networking, green communications, and signal processing.

KHALED B. LETAIEF [S'85, M'86, SM'97, F'03] (eekhaled@ust.hk) received his Ph.D. from Purdue University. He is currently Chair Professor and Dean of Engineering at HKUST. He is an internationally recognized leader in wireless communications with over 500 papers and 15 patents. He is founding Editor-in-Chief of *IEEE Transactions on Wireless Communications* and the recipient of many honors, including the 2009 IEEE Marconi Prize Award in Wireless Communications and 12 IEEE Best Paper Awards. He is an ISI Highly Cited Researcher.

BO BAI [S'09, M'11] (eebobai@tsinghua.edu.cn) received a B.S. (2004) degree with the highest honor from Xidian University (China), and a Ph.D. degree (2010) from Tsinghua University (China). He currently is an assistant professor in the Department of Electronic Engineering, Tsinghua University. He has also obtained the support from the Backbone Talents Supporting Project of Tsinghua University (2012). His research interests include hot topics in wireless communications, information theory, random graph, and combinatorial optimization.

WEI CHEN [S'05, M'07, SM'13] (wchen@tsinghua.edu.cn) received his BS degree in operations research and the Ph.D. degree in electronic engineering (both with the highest honors) from Tsinghua University in 2002, and 2007, respectively. He is a full professor and deputy head of the Electronic Engineering Department, Tsinghua University, as well as a National 973 Youth Project chief scientist and a winner of the National May 1st Medal. He received the IEEE Comsoc APB Best Young Researcher Award and the IEEE Marconi Prize Paper Award.

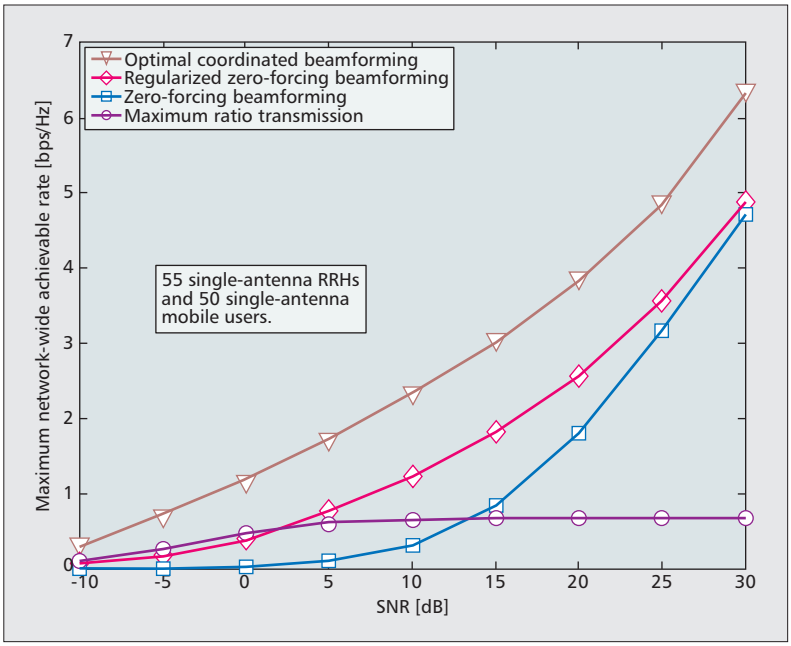

Figure 4. Coordinated beamforming for max-min rate optimization versus transmit signal-to-noise ratio (SNR) in dense Cloud-RAN. The suboptimal transmission strategies only need to perform transmit power optimization as the beamforming directions are pre-fixed.