Recent Advances of Heterogenous Radio Access Networks: A Survey

Yaohua Sun[1] and Mugen Peng[2,*]

[1]*Student Member, IEEE*
[2]*Senior Member, IEEE*
Key Laboratory of Universal Wireless Communications (Ministry of Education), Beijing University of Posts and Telecommunications, Beijing, China
E-mail: sunyaohua@bupt.edu.cn; pmg@bupt.edu.cn
**Corresponding Author*

Received 17 August 2018; Accepted 28 August 2018;
Publication 11 September 2018

Abstract

As a promising paradigm to provide high spectral efficiency and energy efficiency, heterogenous radio access networks (HetNets) have attracted a lot of attention from both academia and industry. This paper presents a comprehensive survey of the recent advances in HetNets, including system architecture evolutions, key techniques, and open issues. The system architectures introduced include conventional HetNets, HetNets with cloud computing, and HetNets with fog computing. In addition, a novel performance metric, together with network self-organization and access slicing, is elaborated, which can help realize a cost-efficient HetNet. Moreover, the other state-of-the-art key techniques in HetNets are surveyed, including non-orthogonal multiple access, interference suppression, and channel estimation in the physical layer, the radio resource allocation in the medium access control layer, and clustering in the network layer. Given the extensiveness of the research area, future research opportunities are identified, which are related to access slicing, HetNets driven by deep learning, and so on.

Journal of Mobile Multimedia, Vol. 14_4, 345–366.
doi: 10.13052/jmm1550-4646.1441

Keywords: HetNets, cost efficiency, signal processing, radio resource allocation, clustering.

1 Introduction

Recent years have witnessed a drastic growth of mobile data traffic, and the heterogenous radio access network (HetNet) has been seen as a promising paradigm to cost-efficiently satisfy such traffic demand [1]. In HetNets, high power nodes (HPNs), such as macro base stations (MBSs), are responsible for wide coverage, while densely deployed low power nodes (LPNs), such as small cell base stations (SBSs), can provide high data rate in hot spots as well as seamless mobility.

However, severe intra-tier and inter-tier interference can occur due to the dense deployment of LPNs, which degrades the performance gains of HetNets, and hence it is essential to adopt some techniques to mitigate the interference. As a key technique in 4G system, the coordinated multiple points (CoMP) is proposed as an effective approach to handle the interference. Nevertheless, its performance is highly constrained by the non-ideal backhaul in real networks. Specifically, the performance of uplink CoMP has been evaluated in downtown Dresden field trials with non-ideal backhaul and distributed cooperative processing located at the base station (BS), and only 20% average spectral efficiency (SE) gains were observed. To further enhance the SE performance and reduce the energy consumption in dense HetNets, the cloud radio access network (C-RAN) is proposed, which incorporates cloud computing technology into RANs, hence facilitating large scale cooperative processing among BSs to effectively suppress interference [2]. Unfortunately, the capacity constrained fronthaul links, which connects the centralized baseband unit (BBU) pool and remote radio heads (RRHs), still have a greatly negative impact on SE and EE gains [3].

As a result, the heterogenous C-RAN (H-CRAN) is advocated in [4], whose main feature is the decoupling of data plane and control plane. Specifically, HPNs equipped with massive MIMO [5] are used to deliver control signalling, while RRHs are used for only data transmission. In this way, the burden of fronthaul links are alleviated, and thus better system performance can be achieved. In addition, the inter-tier interference between RRHs and HPNs can be effectively mitigated by taking advantage of cloud computing capability. Unfortunately, there are several unsolved issues in H-CRANs. First, with video streaming service becoming more and more popular, the fronthaul links suffer many redundant content transmissions. Second, the edge caching and

computing resources are not utilized, which can help alleviate the burden on fronthaul links and the BBU pool significantly.

To overcome the drawbacks of H-CRANs, research on novel RAN architectures should be conducted. As an emerging concept, fog computing refers to the paradigm that endows network edge with storage, communication, control, configuration, measurement, and management capabilities [6]. Motivated by the superior advantages of fog computing, the author in [7] proposes fog radio access networks (F-RANs), which fully utilizes the local cache, local signal processing, and local radio resource management at edge devices like smart user equipments (UEs) to alleviate the burden of fronthaul links. Note that redundant traffic over fronthaul can be effectively avoided by proactively caching popular contents at local cache.

Despite the newly emerging RAN architectures, new performance metrics and other techniques are essential to further enhance the system performance and create a cost-efficient, flexible wireless network with high business value. Given the key role of HetNets in 5G [8] and the recent study achievements, an overview of the architectures, performance metrics, and key techniques of HetNets is presented in this paper. Specifically, RAN architectures include traditional HetNets, H-CRANs, and F-RANs, and a new performance metric called economical energy efficiency, together with access slicing and network self-organization, will be introduced to improve the network cost efficiency. In addition, other key techniques across the physical, MAC, and network layers will be elaborated as well.

The remainder of this paper is organized as follows. Section 2 describes the evolution of HetNet architectures. Section 3 will introduce a novel performance metric named as economical energy efficiency, network self-organization, and access slicing, which can all lead to a more cost-efficient network. Section 4 presents several other key techniques in HetNets, including channel estimation, non-orthogonal multiple access (NOMA), interference suppression, radio resource allocation, and BS clustering. Section 5 identifies several research directions to facilitate future research activities, which are regarded to HetNets driven by deep learning, software-defined-networking (SDN), and so on. Finally, the conclusion is made in Section 6. For convenience, some important abbreviations are listed in Table 1.

2 The Architecture Evolution of HetNets

In this section, various HetNet architectures are introduced, including traditional HetNets, H-CRANs, and F-RANs.

Table 1 Summary of Abbreviations

BS	base station
BBU	baseband unit
CAPEX	capital expenditures
C-RAN	cloud radio access network
CSI	channel state information
D2D	device-to-device
EE	energy efficiency
ESE	economical spectral efficiency
eMBB	enhanced mobile broadband
F-AP	fog access point
F-RAN	fog radio access network
F-UE	fog user equipment
H-CRAN	heterogenous C-RAN
HetNet	heterogenous network
HPN	high power node
HUE	HPN user equipment
IC	interference collaboration
LPN	low power node
MBS	macro base station
mMTC	massive machine-type communications
MSE	mean square error
MUE	macro user equipment
NOMA	non-orthogonal multiple access
OPEX	operational expenditures
QoS	quality of service
RB	resource block
RRH	remote radio heads
RRM	radio resource management
SBS	small cell base station
SDN	software-defined-networking
SE	spectral efficiency
SON	self-organizing networks
UE	user equipment
uMTC	ultra-reliable machine-type communications

2.1 The Architecture of Traditional HetNets

The architecture of a traditional HetNet is illustrated in Figure 1. When subchannels are shared between different tiers, the UEs accessing small cells, such as femtocells, can suffer severe interference from the MBS, and there can also exist intra-tier interference, such as the interference between pico-cells. To

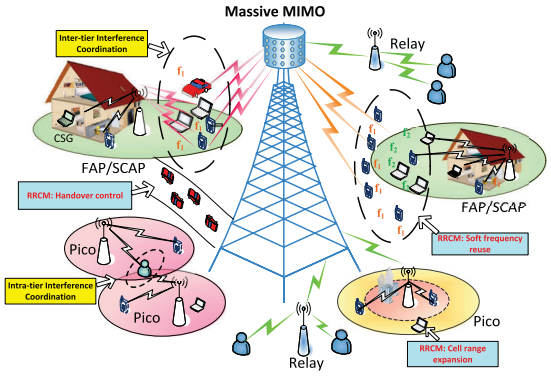

Figure 1 A HetNet Architecture for E-UTRAN systems.

reduce the interference and hence improve system performance, interference coordination and radio resource management (RRM) can be adopted. More concretely, the interference coordination techniques in the physical layer further includes zero-forcing beamforming [9], interference alignment [10], and so on. In addition, by properly allocating power and subchannel resources to different UEs based on their CSI characteristics, RRM can also help mitigate interference. Moreover, handover control and cell range expansion can help better balance the loads among heterogenous nodes, and the deployment of relay nodes can effectively expand the coverage area of the network.

To enhance the performance of cell-edge users, a hierarchical HetNet can be employed, which consists of a hierarchical cooperative basic layer, a homogeneous cooperative enhanced layer, and a heterogeneous cooperative extended layer. In the cooperative basic layer, high-order modulation and coding schemes are utilized to provide cell-edge users with unicast services of high data rate, while hierarchical modulation schemes with unequal error protection space-time code are used for multicasting. In the homogeneous cooperative enhanced layer, user performance is improved by cooperative homogeneous diversity. For the heterogeneous cooperative extended layer, the convergence and interworking of multiple RANs are ensured by heterogeneous cooperative diversity gain.

2.2 The Architecture of H-CRANs

The H-CRAN architecture is illustrated in Figure 2. Similarly to the traditional C-RAN, many RRHs in H-CRANs can perform large scale cooperative

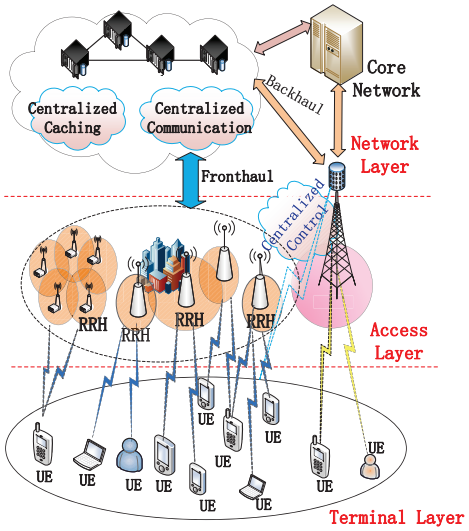

Figure 2 The system architecture of H-CRANs.

transmission owing to the centralized BBU pool, and each RRH possesses low power consumption, since it keeps only the radio frequency and simple symbol processing functionalities. The BBU pool is responsible for executing the remaining physical and upper layer procedures. Nevertheless, there is a big difference between H-CRANs and C-RANs, which lies in the involvement of interfaces between the BBU pool and HPNs. Specifically, S1 and X2 interfaces are defined as the data and control interfaces, respectively, which follows the definitions of 3GPP. Such interfaces facilitate mitigating the inter-tier interference between HPNs and RRHs using centralized cloud computing based cooperative processing techniques. Moreover, both voice and data requirement can be satisfied by H-CRANs, considering the ability of packet switch mode in 4G systems to support voice service. Meanwhile, HPNs control the voice service, while RRHs are deployed to support packet traffic of high data rate.

By moving the delivery of control signallings and system broadcasting information to HPNs, the capacity and latency constraints of fronthaul links can be alleviated, and RRHs can dynamically go into sleep mode according to the network traffic volume, which will decrease the system energy consumption. Besides transmitting control information, HPNs are exploited to support burst traffic and other services with low data rate. Furthermore, a great number of overheads can be saved during radio connection/release

by evolving the pure connection-oriented mechanism to an adaptive one, and HPNs can also be equipped with massive MIMO to enlarge the coverage and raise the capacity. To achieve high data rate for RRH transmission, different techniques in the physical layer can be utilized, including millimeter wave [11], optical light [12], and so on.

2.3 The Architecture of F-RANs

As shown in Figure 3, the F-RAN architecture features fog access points (F-APs) in the access layer and fog UEs (F-UEs) in the terminal layer, which can conduct local caching, local signal processing, and local RRM. For adjacent F-UEs, such as F-UE13 and F-UE11 in Figure 3, they can either communicate directly (device-to-device (D2D) mode) or communicate with the help of another F-UE working as a relay (relay mode). In addition, HPNs and FAPs constitute the access layer, which are responsible for delivering control signalling to F-UEs and forwarding the UE data to the BBU pool over fronthaul links, respectively. Moreover, X2/S1 interface is defined for the backhaul links connecting the BBU pool with HPNs and is backward compatible with that defined in 3GPP standards for LTE and LTE-Advanced systems.

Different from H-CRANs, the burden on the fronthaul and the BBU pool can be greatly reduced by the F-RAN architecture, owing to its capability of processing signal and managing radio resource locally at FAPs and FUEs. Furthermore, by utilizing big data analysis on the user content requests,

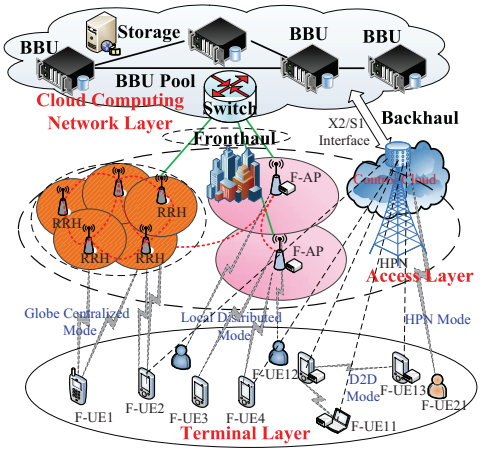

Figure 3 The system architecture for implementing F-RANs.

popular contents can be cached at F-APs and F-UEs in advance, and hence some future requests can be served locally. Considering the popularization of video streaming services, local caching can effectively avoid redundant traffic transmission for these services, and hence alleviate the fronthaul constraints. At last, local caching, local signal processing, and local RRM can better satisfy the low latency requirement of some applications in the 5G era.

3 Towards Cost-Efficient HetNets

Although the technical upgrade has bring users better quality of service (QoS), it is reported that the capital expenditures (CAPEX) and operational expenditures (OPEX) will excess the operator's profits in the future [13]. Hence, in addition to caring about the traditional performance metrics like SE and EE, ever-increasing network cost should also be paid attention to. To this end, a new metric called economical energy efficiency will be introduced, and two techniques, named as self-organizing networks (SON) and access slicing, will be presented, which also both help achieve cost-efficient HetNets.

3.1 Economical Energy Efficiency

Traditional network performance metrics like SE and EE do not take the network cost into account, such as the cost led by the deployment of edge cache and transport networks, which, however, directly impacts the profits of network operators. To evaluate the HetNet in a more comprehensive manner, the author in [14] proposes a novel performance metric, named as economical energy efficiency, which is defined as the ratio of effective system throughput to energy consumption weighted by cost coefficients. Specifically, the weighted energy consumption is a sum of two parts. The first part is the dynamic energy consumption of all the BSs, which relies on the load of each BS, while the second part is the sum of per-BS static power consumption multiplied by a cost coefficient that reflects the heterogeneity of BSs in edge cache resource and X-haul solutions. Numerical results demonstrate that the overuse of X-Haul capacity and edge cache resource has no impact on the SE and EE metrics, while the proposed metric can reflect all the impacts brought by throughput, energy consumption, and affordability.

3.2 Self-Organizing HetNets

As a promising technique, network self-organization endows HetNets with intelligence, which makes it possible to realize the automatical management

and optimization of coverage, capacity, spectrum, and so on, and hence the OPEX of operators can be reduced. In general, the self-organization of HetNets includes the following three aspects: self-configuration, self-optimization, and self-healing [15]. To take full advantage of both centralized and distributed SON architectures, the author in [16] proposes a hybrid SON architecture, and differences between traditional homogeneous networks and HetNets are discussed in terms of self-configuration as well as self-optimization. Since the deployment of a completely self-organized HetNet is non-trivial due to the time-varying traffic, SON coordination in HetNets is essential to be investigated. Specifically, efficient interaction and coordination of SON mechanisms can be realized based on the graph-based decision framework proposed in [17], which describes the interaction and conflict relationships between multiple SON mechanisms using metric event and action graph. Then, for a given event or a combination of events, the strategy-based SON coordination in HetNets can be easily implemented.

3.3 Access Slicing

As a cost-efficient solution to support diverse use cases in the 5G era, network slicing enables the provision of networks in an as-a-service fashion [18]. However, current studies mainly concentrate on core networks based slicing, and the impact of the characteristics of RANs is not well considered. To get a more effective network slicing solution, the author in [19] proposes an enhanced network slicing approach in F-RANs, termed as access slicing. The proposed access slicing architecture is shown in Figure 4, where the networking strategies of three typical scenarios are demonstrated, including massive machine-type communications (mMTC), ultra-reliable MTC (uMTC), and enhanced mobile broadband (eMBB). In the mMTC scenario with massive connections, adjacent UEs are formed into a cluster, and one of the UEs is selected as the cluster head to deliver the traffic of UEs in the cluster to F-APs or HPNs. Moreover, F-APs with caching capability and D2D communication are exploited to achieve low latency in the uMTC scenario, while RRHs cooperatively serve UEs in the eMBB scenario to achieve high data rate.

Meanwhile, the proposed architecture is featured with a centralized orchestration layer and a slice instance layer to guarantee diverse QoS and quality of experience requirements. Specifically, the orchestration layer is responsible for identifying the resource proportion assigned to each slice instance, and the resources can be allocated to slice instances either in an orthogonal

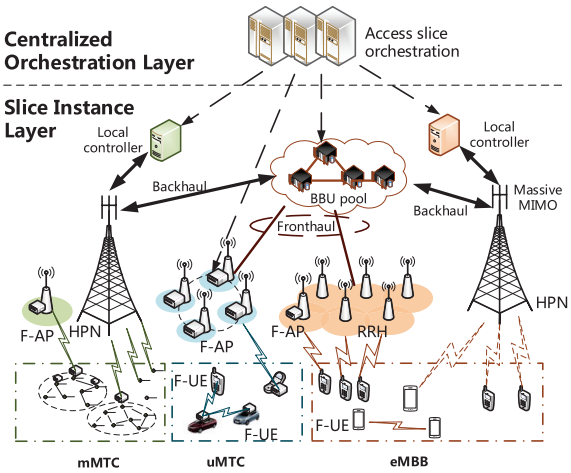

Figure 4 An illustration of access slicing in F-RANs.

or shared manner. In the slice instance layer, each access slice instance controls its own operation and resource allocation. If the dedicated resource can not meet the communication demand, the access slice instance will use the resource shared with other slice instances while limiting the influence on these slice instances.

4 Other Key Techniques

Besides the performance metric and techniques introduced above, other techniques, such as channel estimation in the physical layer and RRM in the MAC layer, are also the key to boosting HetNet performance, which will be presented in this section.

4.1 Key Techniques in the Physical Layer

1) Channel Estimation: In wireless networks, the acquisition of channel state information (CSI) is essential for resource allocation and interference suppression. To achieve a high estimation accuracy while reducing training overhead, the author in [20] proposes a superimposed-segment training design for the uplink of a network with multiple RRHs that connect to the BBU pool via wireless fronthaul, and develops a maximum a posteriori probability channel estimation algorithm. The core idea of the proposal is that each UE superimposes a periodic training sequence on the data signal, and a separate

pilot is prepended to the received signal at each RRH before the signal is forwarded to the BBU pool. By simulation, it is demonstrated that the estimation mean square error (MSE) can be decreased effectively.

Although the proposal in [20] achieves a good performance, there is still room for performance improvement, due to the segment sharing between data and training transmissions. Therefore, the author in [21] studies segment training based channel estimation for an uplink C-RAN. By optimizing the MSE of channel estimation under the power constraint, the training sequence design can be derived for radio access links and wireless fronthaul links. Moreover, to improve the system SE, the lengths of the segment training sequences are optimized.

2) Non-Orthogonal Multiple Access: By enabling multiple users to transmit over the same frequency and time resource at the same time, NOMA can greatly increase the SE and the number of connections for HetNets [22]. In [23], based on stochastic geometry and Gauss-Chebyshev integration, the outage probability is rigourously analyzed for a downlink NOMA-enabled C-RAN with RRHs uniformly distributed in a disk and two users paired to implement NOMA. Via simulations, the correctness of the closed-form expression of the outage probability is verified, and the superiority of the proposal is demonstrated, compared with an opportunistic multiple access scheme and a time division multiple access scheme. While in [24], an NOMA based multicast scheme is proposed for an F-RAN, which promises pushing and multicasting content objects simultaneously, thus leading to a high SE. Using stochastic geometry, the author gives an explicit expression of outage probability for a multi-cell scenario, which shows that the NOMA based multicast scheme outperforms the conventional orthogonal multiple access based multicast scheme.

3) Interference Suppression: In HetNets, when resource blocks are shared among heterogenous nodes, severe inter-tier interference can exist. To suppress the inter-tier interference between RRHs and the HPN in a downlink H-CRAN, two precoding schemes, named as interference collaboration (IC) and beamforming, are investigated for the multi-antenna HPN in [25]. The IC scheme aims at canceling the interference incurred by the transmission of each macro UE (MUE) to all the other MUEs and RRH UEs, while the beamforming scheme intends to only maximize the signal gain for the target UE regardless of the interference to other UEs. By utilizing stochastic geometry, closed-form expressions of the overall outage probabilities, system capacities, and average bit error rates are derived for these two schemes, respectively. Via simulation, it

is concluded that the performance gap between IC and beamforming schemes is related to the number of antennas at the HPN and the number of RRHs.

4.2 Radio Resource Management in the MAC Layer

Effective RRM approaches are essential to fully utilize the limited radio resource in HetNets to meet user QoS. However, resource optimization problems in HetNets are often non-convex and hard to solve due to intra-tier and inter-tier interference, the involvement of binary variables, and so on. Faced with this challenge, authors in [26] and [27] employ various mathematical theories to transform the non-convex problems to more tractable ones. More concretely, in [26], soft fractional frequency reuse is improved for an orthogonal-frequency-division-multiple-access H-CRAN, where RRH UEs with low QoS requirements use the same resource blocks (RBs) with HPN UEs (HUEs). Then, an optimization problem of RB and power allocation is formulated, which aims at maximizing system EE performance. Since the optimization objective is in a fractional form, the formulated problem is non-convex. To handle the non-convexity, the author transforms the primal problem into an equivalent convex feasibility problem, and closed-form resource allocation solutions are derived via Lagrange dual decomposition. While in [27], the secure capacity of D2D pairs is optimized in a D2D underlaying, traditional HetNet. Because of the binary subcarrier allocation variables, the original problem is NP-hard. However, it is observed that the problem under pre-determined subcarrier allocation can be transformed into a convex power allocation problem according to the Perron-Frobenius theory. Therefore, the author is motivated to first derive a sub-optimal subcarrier allocation solution, and then the optimal power allocation problem is solved.

Furthermore, to achieve a better global performance, it is necessary to take into account both the conventional performance metrics in the physical layer and the traffic delay in the upper layer. As a simple framework to deal with delay-aware RRM optimization problems, the Lyapunov optimization approach has been commonly adopted in HetNets. In [28], the author studies the trade-off between throughput utility and delay performance in a slotted H-CRAN. Specifically, a joint stochastic congestion control and resource allocation problem is formulated, which is transformed and decomposed into three separate subproblems that can be concurrently solved at each slot by using the Lyapunov method. Via theoretical analysis, it is shown that the quantitative control of throughput-delay performance trade-off can be achieved with guaranteed EE performance. Instead of taking the EE performance requirement as a constraint in the formulated problem, the author

in [29] aims at maximizing an averaged weighted EE utility objective function directly for H-CRANs under the constraints of inter-tier interference to HUEs and fronthaul capacity constraints. Also leveraging the Lyapunov approach, the primal problem is transformed into a per-slot network-wide cooperative beamformer optimization problem.

In addition to the above works, there is another branch of research that focuses on joint communication mode selection and resource allocation for D2D enabled HetNets. In particular, F-RANs have various ways of communication available to UEs, such as D2D communication mode, HPN mode, global C-RAN mode, and so on, and communication mode selection is generally coupled with resource allocation [30]. With the goal of alleviating the burden on the BBU pool, a distributed reinforcement learning based mode selection and subchannel allocation approach to improving SE is proposed for an uplink F-RAN in [31], which enables each D2D pair to autonomously decide its communication mode and utilized subchannel based on only the received individual utility. The adopted learning algorithm features learning the utility and strategy at the same time based on extremely simple calculations. While in [32], a stochastic optimization problem is investigated under the dynamic CSI and queue state information, whose objective is to maximize the overall average system throughput of F-RANs. Then, this problem is further transformed into a per-slot optimization problem by Lyapunov optimization, which is further decoupled into three sub-problems: the mode selection problem, beamforming design for UEs in global C-RAN mode, and power control for UEs in D2D mode. Based on this decoupling, an iterative algorithm is proposed.

Different from all the above works mainly optimizing traditional performance metrics, a novel performance metric called economical SE (ESE) is maximized for a downlink C-RAN in [33], which takes the energy consumption, fronthaul cost, and system throughput into account. To solve the non-convex ESE maximization problem with fronthaul capacity and transmit power constraints, a beamforming design algorithm containing an outer loop and an inner loop is proposed. In the outer loop, bisection search is utilized to transform the primal problem into an equivalent subproblem, which can be then efficiently solved by a weighted minimum mean square error approach in the inner loop. While in [34], resource allocation for access slicing in an F-RAN is studied, where two kinds of slice instances are considered, namely an eMBB slice and an URLLC slice. By the proposed two-step iterative algorithm that is based on the Hungarian method, linear integer programming, and geometric programming, subcarrier allocation among UEs accessing the

two slices and UE association together with content placement in the URLLC sice are optimized, which minimizes the latency performance for the URLLC slice while satisfying the data rate requirement of the eMBB slice.

4.3 Clustering Schemes in the Network Layer

In wireless networking, it is sometimes beneficial to group BSs into different clusters to attain cooperation or coordination gain within each cluster. To develop a low complexity clustering algorithm, coalitional game based algorithms are often utilized [35]. In [36], an intra-tier interference coordination scheme is proposed for a dense small cell network. Specifically, multiple SBSs can form clusters, within each of which the SBS transmissions are scheduled based on time-division-multiple-access. Then, the interaction among SBSs to form clusters is modeled as a coalition formation game in partition form, aiming at maximizing the individual SBS throughput, and a merge and split based algorithm with partial reversibility rule is developed to achieve a stable partition. In [37], a transfer order based coalition formation algorithm is used to efficiently group FAPs into clusters, and the FAPs in the same cluster perform zero-forcing beamforming cooperatively. The novelty of this game formulation lies in the consideration of cluster formation cost incurred by fetching the missing contents requested by the UEs in the cluster, which can greatly limit the cluster size. Simulation result reveals that increasing the edge caching capacity at FAPs can improve the system SE, since the cost constraint of cluster formation is alleviated, and hence the potential cluster size becomes larger, reducing more intra-tier interference.

5 Future Research Directions

In this section, motivated by the recent trends of wireless networks, several future research opportunities are pointed out.

5.1 HetNets Driven By Deep Learning

As an important technology for enabling artificial intelligence, deep learning [38] has been successfully applied in many areas, including computer vision and speech recognition, and has drawn a lot of attention of researchers in the wireless communication area. In [39], a learning framework is proposed in which convolutional neural networks and recurrent neural networks are used to extract features in spatial domain and time domain from raw information

collected by wireless networks, respectively, and this manner avoids identifying features manually. Taking the extracted features as the state input, deep reinforcement learning is adopted to control wireless networks intelligently, and the superiority of the proposal is verified by applying it to mobility management in wireless local area networks. However, to apply deep learning in HetNets, more studies should be conducted. For example, theoretical analysis should be done to provide guidelines on the architecture design of neural networks and the selection of hyper-parameters. Furthermore, the infrastructure of HetNets needs to be upgraded to support the implementation of deep learning algorithms, especially at the network edge to facilitate real-time network optimization and control.

5.2 Software-Defined Networking and Network Function Virtualization

As two attractive technologies, SDN decouples the control plane from the data plane via controllers [40], while network function virtualization (NFV) aims to decouple network functions from the underlying proprietary hardware [41]. With SDN and NFV, future HetNets can be endowed with high flexibility and reconfigurability, and network functions implemented by softwares can be provisioned on demand, which can both help reduce CAPEX and OPEX and speed up the deployment of new services. In the future, it is interesting to study the practical design of the SDN paradigm for HetNets to achieve intelligent resource management, mobility management, and so on. For the NFV, efficient resource orchestration schemes are expected to better support access slicing in HetNets and meet diverse customization requirements of slice instances.

5.3 Multi-Dimension Resource Allocation

With the convergence of information technology and communication technology, the available resource in future HetNets will further include cache resource and computing resource. Intuitively, to achieve a better global system performance, radio resource, cache resource, and computing resource should be jointly optimized. In [42], the connection between cache resource and radio resource is established by a cache-based adaptive rate requirement. However, such modeling is very simplified, and hence the system model for linking up various resources needs to be further explored. Another big challenge when handling multi-dimension resource allocation is that different kinds of resources can be allocated at different time scales, which makes the problem very complicated [43].

5.4 Access Slicing in HetNets

Although the literature [19] makes a big step for network slicing, several challenges exist for its implementation. First, as stated in [19], the information awareness allows the access slice orchestration layer make intelligent decisions on slice instance creation and resource management. Nevertheless, more specifications should be investigated like the specific information type needed, the frequency for collecting each type of information, and the way of slice instance configuration based on the collected and processed information. Second, resource allocation between multiple slices should be investigated to fully utilize the resources of HetNets to meet diverse performance requirements. However, such a problem can be very challenging, since it is basically a multi-objective optimization problem. Meanwhile, besides allocating radio resource, some slices need cache resource and computation resource, which incurs extra challenges.

5.5 Enabling IoT Services in HetNets

In the 5G era, the IoT services will have more stringent performance requirements, whose scenarios include mMTC and uMTC [44]. The former needs to support massive connections, while the latter should ensure low-latency and highly reliable data transmission. To better support mMTC services, HetNets can be enhanced by incorporating massive MIMO in the physical layer [45] and cluster formation in the network layer [19]. For uMTC services, they can be benefited from HetNets with edge computing capability, such as F-RANs. By moving data analysis and decision-making to the network edge, low latency can be achieved. Nevertheless, there are still many challenges to enable mMTC and uMTC services. For example, although the cluster based access strategy can raise the number of connections, effective and low-complexity cluster formation schemes need to be further investigated. Moreover, computing task allocation algorithms should be developed to assign data analysis or decision making tasks to different fog nodes, taking into account the data availability, computing capability, and communication link condition of each FAP.

6 Conclusion

This paper makes an overview of state-of-the-art system architectures, novel performance metrics, and key technologies of heterogenous networks (HetNets). The architectures surveyed include traditional HetNets, heterogenous

cloud radio access networks, and fog radio access networks, and the economical energy efficiency metric, network self-organization, and access slicing can all help lead to a more cost-efficient HetNet. In addition, other key techniques in the physical, MAC, and network layers, such as non-orthogonal multiple access and radio resource management, are introduced as well. Furthermore, given the extensiveness of the research areas, several future research opportunities are outlined, in terms of the application of deep learning, software-defined-networking, network function virtualization, multi-dimension resource allocation, and so on.

Acknowledgement

This work was supported in part by the State Major Science and Technology Special Project under 2017ZX03001025-006 and 2018ZX03001023-005, the National Natural Science Foundation of China under No. 61831002, and in part by the National Program for Special Support of Eminent Professionals.

References

[1] Peng, M., Wang, C., Li, J., Xiang, H., and Lau, V. K. (2015). Recent advances in underlay heterogeneous networks: Interference control, resource allocation, and self-organization. *IEEE Commun. Surveys Tuts.*, 17(2), 700–729.

[2] Peng, M., Sun, Y., Li, X., Mao, Z., and Wang, C. (2016). Recent advances in cloud radio access networks: System architectures, key techniques, and open issues. *IEEE Commun. Surveys Tuts.*, 18(3), 2282–2308.

[3] Peng, M., Wang, C., Lau, V., and Poor, H. V. (2015). Fronthaul-constrained cloud radio access networks: Insights and challenges. *IEEE Wireless Commun.*, 22(2), 152–160.

[4] Peng, M., Li, Y., Jiang, J., Li, J., and Wang, C. (2014). Heterogeneous cloud radio access networks: A new perspective for enhancing spectral and energy efficiencies. *IEEE Wireless Commun*, 21(6), 126–135.

[5] Larsson, E. G., Edfors, O., Tufvesson, F., and Marzetta, T. L. (2014). Massive MIMO for next generation wireless systems. *IEEE Commun. Mag.*, 52(2), 186–195.

[6] Bonomi, F., Milito, R., Zhu, J., and Addepalli, S. (2012). Fog computing and its role in the internet of things. In *Proceedings of the first edition of the MCC workshop on Mobile cloud computing* (pp. 13–16). ACM.

[7] Peng, M., Yan, S., Zhang, K., and Wang, C. (2016). Fog-computing-based radio access networks: Issues and challenges. *IEEE Netw*, 30(4), 46–53.
[8] Gupta, A., and Jha, R. K. (2015). A survey of 5G network: Architecture and emerging technologies. *IEEE Access*, 3, 1206–1232.
[9] Spencer, Q. H., Swindlehurst, A. L., and Haardt, M. (2004). Zero-forcing methods for downlink spatial multiplexing in multiuser MIMO channels. *IEEE Trans. Signal Process*., 52(2), 461–471.
[10] Cadambe, V. R., and Jafar, S. A. (2008). Interference alignment and degrees of freedom of the K-user interference channel. *IEEE Trans. Inf. Theory*, 54(8), 3425–3441.
[11] Rappaport T. S., *et al.*, (2013). "Millimeter wave mobile communications for 5G cellular: It will work!," *IEEE Access*, 1, 335–349.
[12] Zeng L., *et al.*, (2009). "High data rate multiple input multiple output (MIMO) optical wireless communications using white led lighting," *IEEE J. Sel. Areas Commun.*, 27 (9), 1654–1662.
[13] Checko A., *et al.*, (2015). "Cloud RAN for mobile networks-A technology overview," *IEEE Commun. Surveys & Tutorials*, 17(1), 405–426.
[14] Yan, Z., Peng, M., and Wang, C. (2017). "Economical energy efficiency: An advanced performance metric for 5G systems," *IEEE Wireless Commun.*, 24(1), 32–37.
[15] Imran, A., Zoha, A., and Abu-Dayya, A. (2014). "Challenges in 5G: How to empower SON with big data for enabling 5G," *IEEE Netw.*, 28(6), 27–33.
[16] Peng, M., Liang, D., Wei, Y., Li, J., and Chen, H. (2013). "Self-configuration and self-optimization in LTE-advanced heterogeneous networks," *IEEE Commun. Mag.*, 51(5), 36–45.
[17] Gelabert, X., Sayrac, B., and Jemaa, S. B. (2014). "A heuristic coordination framework for self-optimizing mechanisms in LTE HetNets," *IEEE Trans. Veh. Technol.*, 63(3), 1320–1334.
[18] Samdanis, K., Costa-Perez, X., and Sciancalepore, V. (2016). "From network sharing to multi-tenancy: The 5G network slice broker," *IEEE Commun. Mag.*, 54(7), 32–39.
[19] Xiang, H., Zhou, W., Daneshmand, M., and Peng, M. (2017). "Network slicing in fog radio access networks: Issues and challenges," *IEEE Commun. Mag.*, 55(12), 110–116.
[20] Xie, X., Peng, M., Wang, W., and Poor, H. V. (2015). "Training design and channel estimation in uplink cloud radio access networks," *IEEE Signal Process. Lett.*, 22(8), 1060–1064.

[21] Hu, Q., Peng, M., Mao, Z., Xie, X., and Poor, H. V. (2016). "Training design for channel estimation in uplink cloud radio access networks," *IEEE Trans. Signal Process.*, 64 (13), 3324–3337.

[22] Ding, Z., Peng, M., and Poor, H. V. (2015). "Cooperative non-orthogonal multiple access in 5G systems," *IEEE Commun. Lett.*, 19(8), 1462–1465.

[23] Gu, X., Ji, X., Ding, Z., Wu, W., and Peng, M. (2018). "Outage probability analysis of non-orthogonal multiple access in cloud radio access networks," *IEEE Commun. Lett.*, 22(1), 149–152.

[24] Zhao, Z., Xu, M., Li, Y., and Peng, M. (2017). "A non-orthogonal multiple access based multicast scheme in wireless content caching networks," *IEEE J. Sel. Areas Commun.*, 35(12), 2723–2735.

[25] Peng, M., Xiang, H., Cheng, Y., Yan, S., and Poor, H. V. (2015). "Inter-tier interference suppression in heterogeneous cloud radio access networks," *IEEE Access*, 3, 2441–2455.

[26] Peng, M., Zhang, K., Jiang, J. Wang, J., and Wang, W. (2015). "Energy-efficient resource assignment and power allocation in heterogeneous cloud radio access networks," *IEEE Trans. Veh. Technol.*, 64(11), 5275–5287.

[27] Zhang, K., Peng, M., Zhang, P., and Li, X. (2017). "Energy-efficient resource assignment and power allocation in heterogeneous cloud radio access networks," *IEEE Trans. Veh. Technol.*, 66(2), 1822–1834.

[28] Li, J., Peng, M., Yu, Y., and Ding, Z. (2016). "Energy-efficient joint congestion control and resource optimization in heterogeneous cloud radio access networks," *IEEE Trans. Veh. Technol.*, 65(12), 9873–9887.

[29] Peng, M., Yu, Y., Xiang, H., and Poor, H. V. (2016). "Energy-efficient resource allocation optimization for multimedia heterogeneous cloud radio access networks," *IEEE Trans. Multimedia*, 18(5), 879–892.

[30] Gao C., *et al.*, (2016). "Enabling green wireless networking with device-to-device links: A joint optimization approach," *IEEE Trans. Wireless Commun.*, 15(4), 2770–2779.

[31] Sun, Y., Peng, M., and Poor, H. V. (2018). "A distributed approach to improving spectral efficiency in uplink device-to-device enabled cloud radio access networks," *IEEE Trans. Commun.*, doi: 10.1109/TCOMM.2018.2855212, submitted for publication.

[32] Mo, Y., Peng, M., Xiang, H., Sun, Y., and Ji, X. "Resource allocation in cloud radio access networks with device-to-device communications," *IEEE Access*, 5, 1250–1262.

[33] Peng, M., Wang, Y., Dang, T., and Yan, Z. (2017). "Cost-efficient resource allocation in cloud radio access networks with heterogeneous fronthaul expenditures," *IEEE Trans. Wireless Commun.*, 16(7), 4626–4638.
[34] Tang, L., Zhang, X., Xiang, H., Sun, Y., and Peng, M. (2017). "Joint resource allocation and caching placement for network slicing in fog radio access networks," in *Proceedings of SPAWC*, Sapporo, Japan, 1–6.
[35] Pantisano, F., Bennis, M., Saad, W., Debbah, M., and Latva-aho, M. (2013). "Interference alignment for cooperative femtocell networks: A game-theoretic approach," *IEEE Trans. Mobile Comput.*, 12(11), 2233–2246.
[36] Ahmed, M., Peng, M., Abana, M., Yan, S., and Wang, C. (2018). "Interference coordination in heterogeneous small-cell networks: A coalition formation game approach," *IEEE Syst. J.*, 12(1), 604–615.
[37] Sun, Y., Dang, T., and Zhou, J. (2016). "User scheduling and cluster formation in fog computing based radio access networks," in *Proceedings of ICUWB*, Nanjing, China, 1–4.
[38] LeCun, Y., Bengio, Y., and Hinton, G. (2015). "Deep learning," *Nature*.
[39] Cao, G., Lu, Z., Wen, X., Lei, T., and Hu, Z. (2018). "AIF: An artificial intelligence framework for smart wireless network management," *IEEE Commun. Lett.*, 22(2), 400–403.
[40] Sezer, S., *et al.*, (2013). "Are we ready for SDN? Implementation challenges for software-defined networks," *IEEE Commun. Mag.*, 51(7), 36–43.
[41] Hawilo, H., *et al.*, (2014). "NFV: State of the art, challenges, and implementation in next generation mobile networks," *IEEE Netw.*, 28(6), 18–26.
[42] Tang, J., Teng, L., Quek, T. Q. S., Chang, T., and Shim, B. (2017). "Exploring the interactions of communication, computing and caching in cloud RAN under two timescale," in *Proceedings of SPAWC*, Sapporo, Japan, 1–6.
[43] Tang, J., Wen, R., Quek, T. Q. S., and Peng, M. (2017). "Fully exploiting cloud computing to achieve a green and flexible C-RAN," *IEEE Commun. Mag.*, 55(11), 40–46.
[44] Bockelmann, C., *et al.*, (2016). "Massive machine-type communications in 5G: Physical and MAC-layer solutions," *IEEE Commun. Mag.*, 54(9), 59–65.
[45] Liu, L., and Yu, W. (2018). "Massive connectivity with massive MIMO-Part I: Device activity detection and channel estimation," *IEEE Trans. Signal Process.*, 66(11), 2933–2946.

Biographies

Yaohua Sun (S'18) received the bachelor's degree (with first class Hons.) in telecommunications engineering (with management) from Beijing University of Posts and Telecommunications (BUPT), Beijing, China, in 2014. He is a Ph.D. student at the Key Laboratory of Universal Wireless Communications (Ministry of Education), BUPT. His research interests include game theory, resource management, deep reinforcement learning, network slicing, and fog radio access networks. He was the recipient of the National Scholarship in 2011 and 2017, and he has been reviewers for *IEEE Transactions on Communications*, *Journal on Selected Areas in Communications*, *IEEE Communications Magazine*, *IEEE Communications Letters*, and *IEEE Internet of Things Journal*.

Mugen Peng (M'05, SM'11) received the Ph.D. degree in communication and information systems from the Beijing University of Posts and Telecommunications (BUPT), Beijing, China, in 2005. Afterward, he joined BUPT, where he has been a Full Professor with the School of Information and Communication Engineering since 2012. During 2014 he was also an academic visiting fellow at Princeton University, USA. He leads a Research Group focusing on wireless

transmission and networking technologies in BUPT. He has authored and co-authored over 90 refereed IEEE journal papers and over 300 conference proceeding papers. His main research areas include wireless communication theory, radio signal processing, cooperative communication, self-organization networking, heterogeneous networking, cloud communication, and Internet of Things.

Dr. Peng was a recipient of the 2018 Heinrich Hertz Prize Paper Award, the 2014 IEEE ComSoc AP Outstanding Young Researcher Award, and the Best Paper Award in the JCN 2016, IEEE WCNC 2015, IEEE GameNets 2014, IEEE CIT 2014, ICCTA 2011, IC-BNMT 2010, and IET CCWMC 2009. He is currently or have been on the Editorial/Associate Editorial Board of the *IEEE Communications Magazine*, *IEEE ACCESS*, *IEEE Internet of Things Journal*, *IET Communications*, and *China Communications*. He received the First Grade Award of Technological Invention Award three times in China.