ULTRA-LOW-LATENCY UBIQUITOUS CONNECTIONS IN HETEROGENEOUS CLOUD RADIO ACCESS NETWORKS

SHAO-YU LIEN, SHAO-CHOU HUNG, KWANG-CHENG CHEN, AND YING-CHANG LIANG

Shao-Yu Lien is with National Formosa University.

Shao-Chou Hung and Kwang-Cheng Chen are with National Taiwan University.

Ying-Chang Liang is with the Institute for Information Research.

ABSTRACT

Although heterogeneous cloud radio access networks (H-CRAN) have emerged for efficient network management, resource management, and throughput enhancement, new technical challenges remain to support the urgent need to achieve full automation for ultra-low-latency connections. To simultaneously accommodate such diverse technical requirements in the H-CRAN, in this article, a new paradigm of H-CRAN design to significantly reduce latency is presented. From the refined radio access in the air interface, new methodology of radio resource optimization in the system architecture, and intelligent routing/paging in the backhaul, the provided foundation suggests open-loop communications and other new directions to facilitate state-of-the-art practices of the H-CRAN.

INTRODUCTION

Achieving full automation to largely enhance human beings' sensory and processing capabilities has been regarded as one of the ultimate goals in recent network services. This goal embraces all emerging applications such as unmanned or remotely controlled robots/vehicles/offices/factories, augmented/virtual reality, intelligent transportation systems, smart grid/building/city, immersive sensory experience, and the Internet of Things (IoT). Very different from conventional multimedia and file delivery services, there are two major characteristics of these emerging applications.

- Control, data collection, and sensing environments play a considerable role in these applications, which only generate packets with a small size (i.e., few coded bytes). However, as the number of devices involved in these applications can be tremendous, the amount of small packets can be extremely large. To support such massive small packet transmissions, the spectrum efficiency and complexity of network management are very critical concerns. Wide deployment of IoT will make this issue even more challenging.
- These applications are extremely vulnerable to data transmission latency. Providing ultra-low-latency (at the millisecond level) services consequently turns out to be the most crucial requirement. However, the state-of-the-art fourth generation (4G) mobile network cannot support the needs toward this goal. The reason comes from the fact that the major requirement of the 4G mobile network is to considerably boost communication data rates, and thus sophisticated radio procedures such as channel estimation, link adaptations, channel-aware scheduling, and resource allocation are adopted to make good use of all possible radio resources in the air interface. To support these procedures, heavy signaling overheads are imposed on both the wired and wireless links, which eventually affect the energy efficiency and latency performance needed to support full automation. To provide ultra-low-latency services, the latency performance of existing mobile networks needs to be improved 100- to 1000-fold. Without new designs and fundamental inspirations, this unprecedented goal may not be fulfilled by the existing mobile networks.

To support the first characteristic, enhanced spectrum efficiency, a remarkable technology known as heterogeneous networks (HetNets) [1] has been adopted by the existing mobile networks. By deploying small cells (e.g., picocells, femtocells, WiFi offloading, or relay nodes) or allowing devices to directly exchange data (known as device-to-device, D2D) overlaying existing macrocells, HetNets have been shown as a key enabler to enhance spectrum reuse and the signal-to-interference-plus-noise ratio (SINR) for each communication link. In LTE/LTE-A HetNet, each base station, or eNB, of a macrocell/picocell/femtocell/relay node is able to autonomously allocate/schedule radio resources for its mobile devices. However, individually managing radio resources in each eNB may invoke severe intercell interference. This defect consequently motivates the idea of joint resource allocation among eNBs. Through the provided wired/wireless interfaces (i.e., S1, X2, and Un interfaces) among eNBs, remote radio heads (RRHs), baseband units (BBUs), multiple eNBs,

1536-1284/15/$25.00 © 2015 IEEE

RRHs/BBUs, and relay nodes are also able to exchange information for joint resource scheduling/allocation. Under this paradigm, the promising cloud computing technology can be applied to facilitate the optimization of joint resource allocation. This architecture is known as the heterogeneous cloud radio access network (H-CRAN) first revealed in [2–4]. In the H-CRAN, a set of radio data and computing servers collect channel state information (CSI), interference levels, and quality of service (QoS) requirements in all cells to achieve the global optimum resource allocation, as shown in Fig. 1. In other words, multiple cells in the H-CRAN are transparent to mobile devices as a single "big" cell.

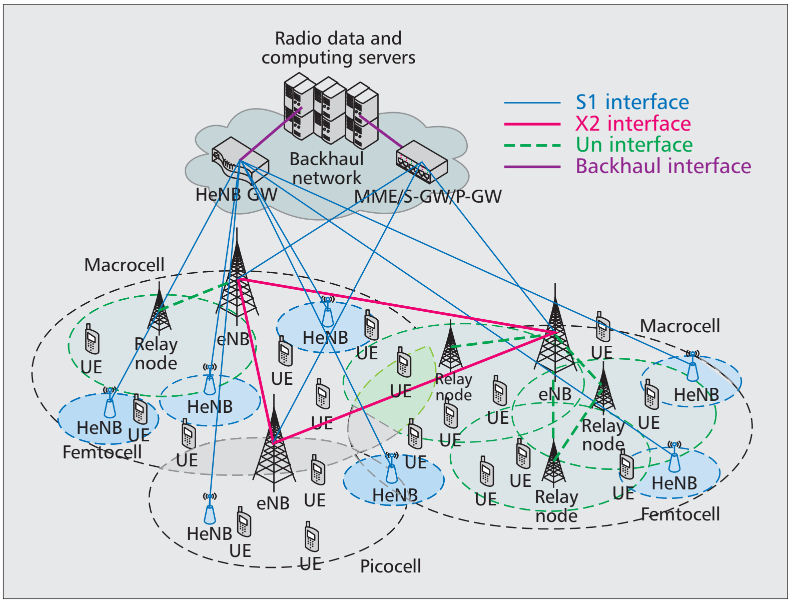


Figure 1. In the H-CRAN, eNBs of macrocell/picocell/femtocell and relay nodes are able to forward channel state information and interference levels in each cell to a set of radio data and computing servers via S1, X2, and Un interfaces. The servers perform a centralized optimization of resource allocation for each cells. Multiple "cells" in the H-CRAN are thus transparent to mobile devices as a single "big" cell.

Although the H-CRAN has shown the potential in spectrum efficiency and energy efficiency enhancements [2, 3], to support emerging applications, the second characteristic of providing ultra-low latency services is still a huge challenge. To enhance the end-to-end latency performance, a proper resource scheduling/allocation scheme is able to reduce the transmission delay in the air interface [5]. Adopting an access control policy is also effective to reduce latency in wired/wireless backhaul [6]. However, the unacceptable signaling overheads in the air interface still impose large data exchange delay on existing mobile networks. In addition to latency resulting from signaling overheads, the H-CRAN further induces two new sorts of latency which may be more severe than that in the air interface. The first one is latency in resource optimization. Radio resource optimization is a widely discussed issue, and the existing results reveal that the computational complexity is increased along with the number of available radio resources, the number of devices, and the number of eNBs. Unfortunately, it is projected that the number of devices will exponentially increase in the following decade. To support the increasing number of devices, the number of available radio resources should be increased as well as the number of eNBs in the H-CRAN. This growing computational complexity may eventually obstruct the latency performance. The second one is latency in the routing/paging procedures to forward data to a mobile device in the H-CRAN. In the existing mobile network design, the routing and paging procedures assume that each mobile device may communicate with any other mobile devices and servers. Therefore, fixed routing/paging information with a hierarchical information inquiry scheme is adopted. However, this design fully ignores the fact that a mobile device may frequently communicate with the mobile devices or web servers within its social network, while rarely exchanging data with terminals/servers outside its social network. The existing routing/paging procedures may therefore result in a more severe delay than that in the air interface.

To achieve millisecond-level latency over the H-CRAN, three types of latency shall be alleviated in the H-CRAN:

- Latency in the radio access
- Latency in the optimization computation
- Latency in routing and paging

In the article, we consequently reveal a series of mechanisms and principles. From the open-loop radio access at the link level to alleviate signaling overheads, information-bridled resource optimization at the system level to reduce complexity, to the social data cache-based architecture at the network level to diminish routing/paging latency, a complete design philosophy as well as performance evaluation are provided as essential foundations for the next generation H-CRAN.

Open-Loop Radio Access

Reviewing the history of the development of mobile networks from the third generation (3G) network (Universal Mobile Telecommunications System, UMTS) to the 4G network (Long Term Evolution/Advanced, LTE/LTE-A), closed-loop communications are adopted to ubiquitously track available radio resources. In closed-loop communications, the optimum (downlink/uplink) data transmission schemes (i.e., dynamic link adaptations and resource scheduling/allocation) for a transmitter are decided after (uplink/downlink) feedback information from the receiver is obtained. For example, the inner-loop power control performs 1500 times per second in UMTS, while the channel estimation performs up to 1000 times per second in LTE/LTE-A. In addition, automatic repeat request (ARQ) and hybrid ARQ (HARQ) are also adopted in LTE/LTE-A, which need acknowledgment messages sent from a receiver to indicate to the transmitter successful data reception. Although the closed-loop operation in 3G and 4G mobile networks provides reliable and high data rate communications for streaming multimedia traffic, it is disfavored and challenging to support the following scenarios.

Closed-loop invokes poor spectrum efficiency

For reliable communications, a transmitter thus needs to autonomously determine the transmission strategy, which includes the modulation and coding scheme, the space-time code for MIMO communications, and the number of transmission repetitions in the time, frequency, and spatial domains in one-shot.

and energy efficiency in massive uplink transmissions. To achieve full automation, there can be trillions of mobile devices with a significant amount of small packets. In uplink transmissions, when a massive amount of small packets are uploaded to the H-CRAN, the H-CRAN needs to acknowledge all packets in the downlink channel. As a result, the downlink channel may be blocked, subsequently terminating all communications. In addition, transmitting massive acknowledgment messages also significantly harms the energy consumption at the H-CRAN.

Closed-loop harms the latency performance. In closed-loop communications, a receiver requests a transmitter to retransmit a packet if this packet cannot be correctly received (i.e., ARQ or HARQ). Further assessing the risk of the reception failure of acknowledgment messages, a link failure is eventually identified when the maximum allowable retransmission number is reached. The latency may not be acceptable for full automation. In addition, for the case of multihop (relay) transmissions (e.g., relay transmissions in Third Generation Partnership Project, 3GPP, Rel-10/Rel-11 and D2D relay transmissions in 3GPP Rel-14), the latency at each hop is accumulated, making closed-loop communications infeasible.

The Potential of Open-Loop Radio Access

To considerably reduce the signaling overheads, in 3GPP Rel-12 and Rel-13, feedback reduction designs are adopted for low-cost machine-type communication (MTC) devices. By relaxing the number of feedback messages, it is anticipated that the energy consumption and transmission latency for MTC devices transmitting very small packets will be reduced. However, relaxing the amount of feedback messages also suggests the lack of full CSI; therefore, MTC devices only adopt conservative link adaptation schemes to avoid 64-quadrature amplitude modulation (QAM) and multiple-input multiple-output (MIMO) communications. This state-of-the-art paradigm motivates us to rethink the system architecture based on open-loop communications. In open-loop communications, a receiver does not provide feedback information (e.g., CSI and acknowledgment messages) to a transmitter. Without feedback information, a transmitter has no knowledge of the channel condition as well as successful reception at the receiver side. For reliable communications, a transmitter thus needs to autonomously determine the transmission strategy, which includes the modulation and coding scheme, the space-time code for MIMO communications, and the number of transmission repetitions in the time, frequency, and spatial domains in one shot.

In conventional closed-loop communications, a link failure can be identified by both the transmitter and the receiver until the maximum retransmission limit of ARQ/HARQ has been reached. It therefore needs an amount of transmission redundancy (retransmissions) and a period of waiting time. On the contrary, for open-loop communications, the major concern lies in the capability of providing reliable communications. To protect critical data, a transmitter can adopt a very conservative modulation and coding scheme or a large amount of redundant resources for transmission repetitions in one shot. It therefore only needs an amount of transmission redundancy to avoid a period of waiting time to identify a link failure. Please note that providing error-free communications is practically impossible in both open-loop and closed-loop communications, due to the fact that the channel variation and noise effect are stochastic. If the channel fading is not severe, both open-loop and closed-loop communications have opportunities to correctly receive data. On the other hand, if the channel fading is indeed severe (i.e., a link failure occurs), reliable communications are infeasible for both open-loop and closed-loop communications. Nevertheless, a link failure can be quickly identified in open-loop communications to significantly reduce latency and obtain more time for further processes by upper layers.

The technical merits of open-loop communications can be fully revealed in Fig. 2. In Fig. 2a, seven small cells (i.e., HeNBs of femtocells) and a number of devices are randomly deployed. Each device attaches to the HeNB with the highest signal strength. However, considering the uplink transmissions, as the density of devices increases, the SINR for each uplink transmission at the HeNB may decrease due to increasing interference. To combat interference, in closed-loop communications, HARQ and different levels of signaling overheads are needed. In Fig. 2a, 10, 20, and 30 percent signaling overheads (compared to the amount of transmitted data) are considered. These overheads include traffic of acknowledgment messages and channel estimation reports, and the HARQ retransmission number limit is set to 10. In closed-loop communications, when an error occurs in a data transmission, this data transmission will be repeated after the notification of acknowledgment messages. If the density of devices is high (and thus the SINR is low), retransmissions may be performed several times, which severely harms the latency performance. On the other hand, in open-loop communications, to combat interference, redundant resources (for a conservative modulation and coding scheme and transmission repetitions) are needed, while the amount of these redundant resources is decided in one shot (with 2-, 4-, and 6-fold redundancy compared to the amount of transmitted data). Therefore, we can observe from Fig. 2a that to achieve a successful transmission, the latency performance of open-loop communications is significantly enhanced compared to that of closed-loop communications when the density of devices increases. Since redundant resources are required in both closed-loop and open-loop communications, whether the spectrum efficiency of open-loop communications may be worse than that of closed-loop communications needs to be investigated. We can observe from Fig. 2b that as the density of devices increases, the spectrum efficiency of closed-loop communications decays dramatically. On the other hand, the spectrum efficiency of open-loop communications is competitive as the density of devices increases. This result demonstrates that open-loop communications are practical for the H-CRAN. We also evaluate the latency performance of multihop and multi-path transmissions in the H-CRAN to support multimedia traffic in Fig. 2c. In multihop relay transmissions, if a link failure occurs at a hop, the source node may fail to deliver a packet to the destination node on time. To enhance the latency performance, the source node may replicate (or network code) the packet transmission simultaneously via multiple paths [7]. In this case, the latency is unacceptable only if the source node fails to deliver a packet to the destination on time at all paths. As a result, there is a trade-off between latency and the number of redundant paths. In Fig. 2c, 50 disjoint paths are deployed, and each path is composed of a number of links (hops) randomly selected from [1, 5]. In closed-loop communications, HARQ is applied to each hop transmission, and therefore latency at all hops in a path is accumulated. We can observe from Fig. 2c that closed-loop communications requires more redundant paths than open-loop communications to support five voice (VoIP) and five video (MPEG4) streams. This result confirms the latency performance improvement via the open-loop communications, especially in multihop relay transmissions.

Cognitive Radio Technology and Cognitive Radio Resource Management

In open-loop communications, the transmission scheme is autonomously determined by a transmitter. To optimize the transmission scheme without any knowledge at the receiver side, each transmitter thus needs to infer CSI, interference levels, and even the transmission scheme adopted by other transmitters. This operation is very similar to cognitive radio (CR) technology. In the scenario of CR networks [8], secondary users opportunistically access the channel when the primary users are absent. In such opportunistic channel access, the reverse link of secondary users does not always exist. Each transmitter of secondary users thus autonomously senses the communication environment to avoid interference from primary users as well as from transmitters of other secondary users. This similar radio behavior makes CR technology completely compatible with open-loop radio access. The CR technology can be viewed as a sort of connectionless transmission scheme in the air interface, which facilitates significant overhead savings for burst traffic with a small amount of data [2] by avoiding a pure connection-oriented mechanism of closed-loop communications. The CR technology can be implemented in different forms for devices with different capabilities. For low-cost (MTC) devices, the hardware and computing competence are limited; hence, sophisticated CR functions may not be able to be supported. In this case, the CR technology can be realized as simply the listen-before-talk scheme with energy-detection-based clear channel assessment to detect the presence of interference. If interference is detected, a transmitter is able to transmit data with a large amount of redundancy to protect data, or suspend transmissions to avoid interference. For devices with mighty capabilities, the CR technology can be realized to be

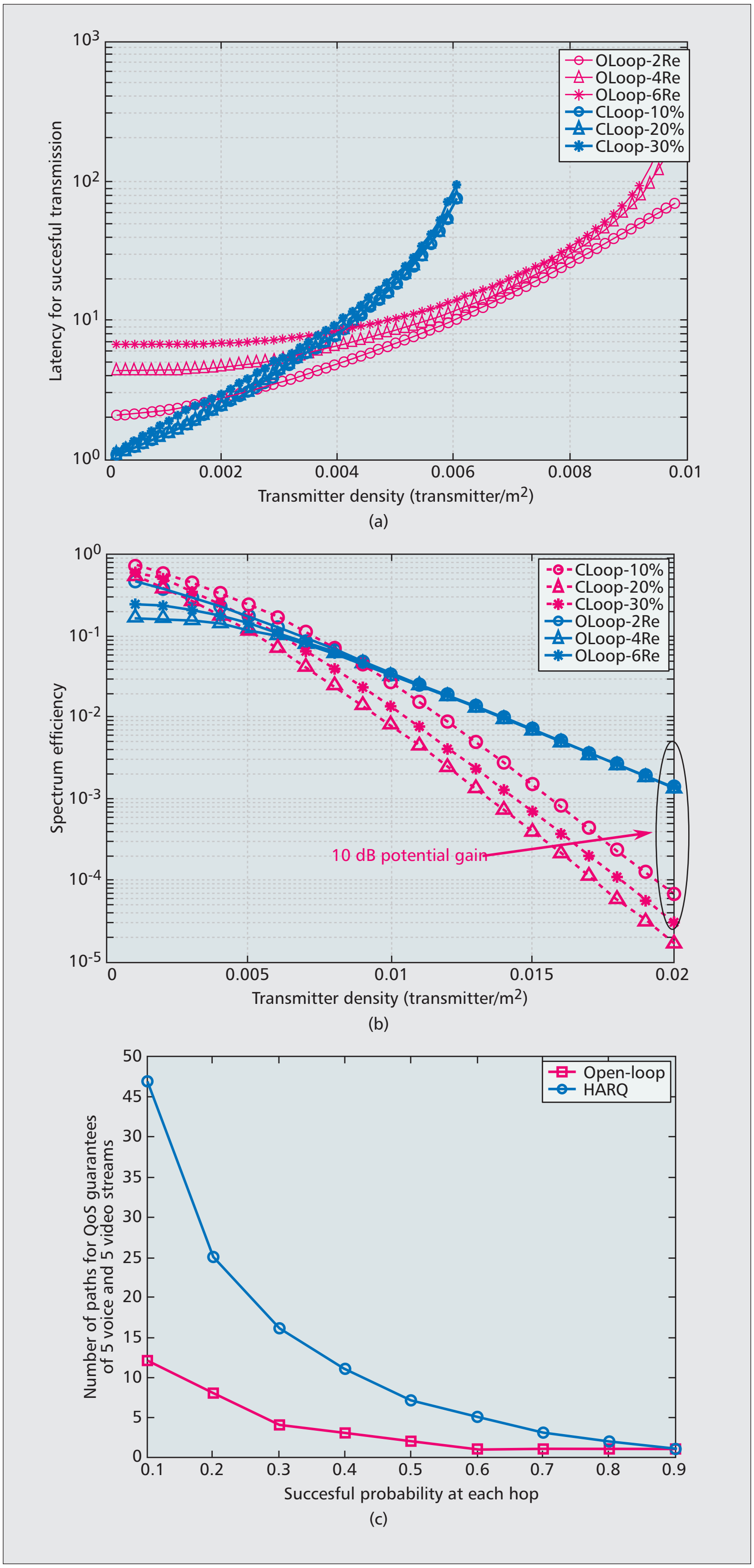


Figure 2. a) Latency performance for successful (error-free) transmission of open-loop and closed-loop communications under different densities of deployed devices; b) spectrum efficiency of open-loop and closed-loop communications under different densities of deployed devices; c) number of redundant paths needed to provide QoS guarantees of five VoIP and five MPEG4 streams under different levels of channel fading (in terms of successful probability) at each hop. In this simulation, the number of hops in each path is randomly selected from [1, 5].

As the numbers of mobile devices, packets, available resources, and QoS requirements increase, the complexity of a central optimization also significantly grows. The latency of resource optimization computation thus eventually becomes the performance bottleneck.

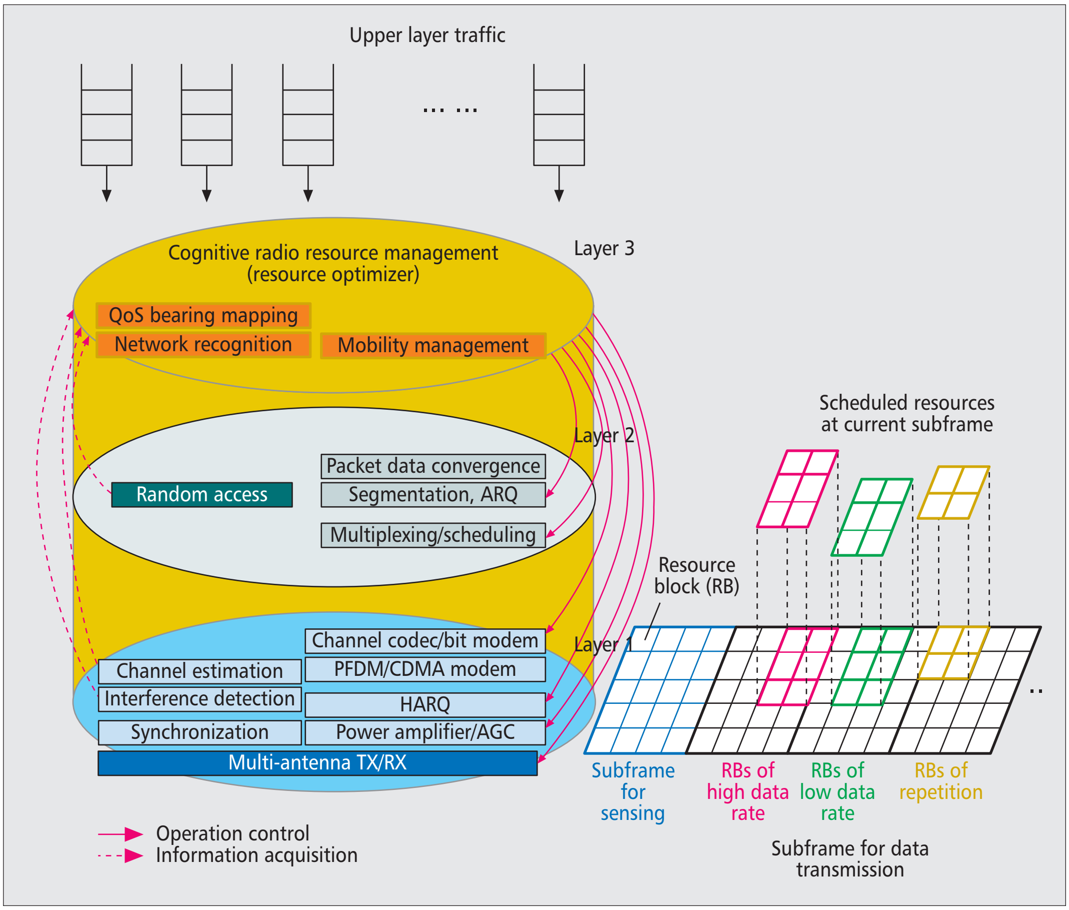


Figure 3. The laylerless design of CRRM for open-loop communications.

intelligent to optimize the overall network performance as well as a user's particular needs.

Recently, in [9], cognitive radio resource management (CRRM) is demonstrated to apply CR to mitigate interference in the HetNet. The CRRM is a top-down radio design, by which the radio resource control layer of a transmitter jointly optimizes layer 1 and layer 2 resource allocations for channel sensing, data transmissions, and interference avoidance by taking the upper layer QoS requirements (bearing establishment) into account, as illustrated in Fig. 3. Applied to the LTE/LTE-A system, the CRRM periodically allocates a subframe of radio resources to perform channel measurement for interference detection (this subframe is known as a measurement subframe). If interference is detected on certain radio resources, these radio resources are not utilized for data transmissions in subsequent subframes (these subframes are known as data subframes). Frequently allocating measurement subframes can fully capture channel variation as well as interference. However, measurement subframes are a sort of overhead, as data transmissions cannot be performed within a measurement subframe in order to detect interference. A proper period of allocating measurement subframes is decided based on the latency requirement of upper layer traffic. As a result, CRRM converts the existing protocol stacks of mobile networks into a sort of layerless design. Such a design is particularly crucial for small packet transmissions, reducing latency of packet and protocol conversion between layers, alleviating overheads and processing delay imposed to packets at each layer, and making the radio behavior more flexible. Due to the coherence of the design goal, CRRM is sufficiently compatible with the open-loop radio access for each transmitter to autonomously decide on a transmission scheme that also supports traditional broadband multimedia services.

Information-Bridled Resource Optimization in the H-CRAN

After alleviating latency in the air interface, the next challenge in the H-CRAN lies in scalability. As the numbers of mobile devices, packets, available resources, and QoS requirements increase, the complexity of a central optimization also significantly grows. The latency of resource optimization computation thus eventually becomes the performance bottleneck. Although distributed resource allocation at each eNB may reduce the complexity, the performance is degraded. To tackle this challenge, let us rethink the foundation of optimization theory: that the performance of an optimization is subject to available information. In the H-CRAN, the computing servers collect all radio information in each cell for centralized optimization. If all radio information in each cell is available for each mobile device, individual resource optimization performed at each device is able to achieve the same performance as that of the centralized optimization. However, this scheme is practically infeasible since exchanging radio information among all devices induces unacceptable signaling overheads. Nevertheless, we may resolve this issue via open-loop communications.

To provide reliability in open-loop communications, a transmitter does not need instantaneous CSI. Instead, a transmitter needs the long-term statistics of CSI, interference levels, and transmission schemes adopted by other transmitters. Such radio information can be socially obtained through CR technology, or globally provided by the network. Given radio information, each transmitter is thus able to optimize its transmission scheme. In this paradigm, the computing servers in the H-CRAN do not have to conduct centralized resource optimization. Instead, the H-CRAN only needs to tailor radio information provided to all transmitters. As the optimization procedure performed by each transmitter with given radio information is well known by the H-CRAN, the H-CRAN can control all transmitters through controlling/optimizing information fed to transmitters. This concept leads to a new architecture of information-bridled resource optimization in the H-CRAN, as illustrated in Fig. 4. In this architecture, *the H-CRAN only optimizes radio information provided to transmitters; then the radio accesses of all transmitters are under the control of the H-CRAN, as all transmitters optimize their transmission schemes based on provided radio information*. The principles of the information-bridled resource optimization are summarized as follows, in which, we use the uplink transmission as an elaboration example.

- The H-CRAN broadcasts a set of radio information $\mathbf{R} = [r_1, r_2, \ldots, r_K]$ regarding each radio resource for all mobile devices (transmitters), where K is the number of total radio resources in a scheduling period. By taking the statistics of CSI, interference levels, and transmission schemes adopted by all mobile devices into account, radio information for the kth radio resource is mapped into a number r_k, which is normalized to $0 \leq r_k \leq 1$. Radio resources can be defined in different domains, such as time, frequency, eNB, code, or spatial domain. r_k of a lower value reveals that the kth radio resource suffers from severe channel fading, interference, or congestion. It distracts mobile devices from selecting the kth radio resources. On the other hand, r_k of a higher value implies that the kth radio resource enjoys better channel quality, lower interference, or mild congestion. Such indication attracts mobile devices to select the kth radio resource.
- Upon receiving a set of radio information $\mathbf{R} = [r_1, r_2, \ldots, r_K]$, each mobile device autonomously selects a radio resource by taking the resource selection strategies adopted by other mobile devices into consideration. That is, a mobile device should not always select the radio resource with the highest r_k, since other mobile devices may also select this radio resource, leading to severe interference and congestion. Game theory can be an effective means for all mobile devices to determine the transmission scheme in this scenario to optimize utility. After selecting a radio resource, each mobile device then transmits data to the H-CRAN via this radio resource in the open-loop fashion.
- After the transmissions from all mobile devices, the H-CRAN are able to obtain interference and congestion levels at all radio resources. In addition, the H-CRAN needs to estimate CSI of all radio resources at all locations (known as the spectrum map). The spectrum map can be estimated via the uplink transmissions from all mobile devices [10]. Then the H-CRAN optimizes/updates $\mathbf{R} = [r_1, r_2, \ldots, r_K]$ and broadcasts this radio information.

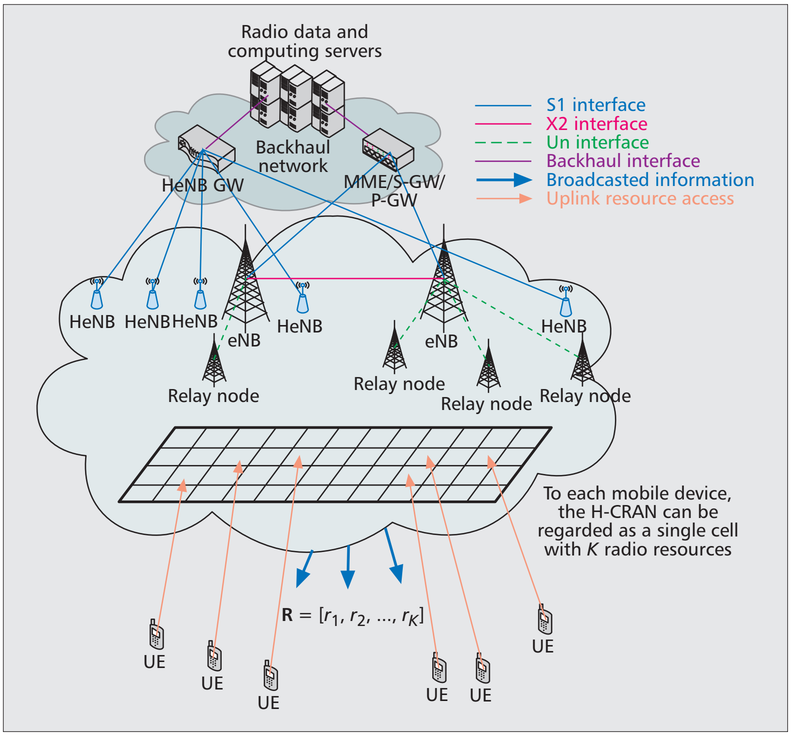


Figure 4. Information-bridled resource optimization, in which the H-CRAN only optimizes radio information broadcast to all mobile devices. Given radio information, each mobile device autonomously optimizes the transmission scheme. Thus, the H-CRAN controls the radio accesses of all mobile devices implicitly.

As aforementioned, the complexity of conventional resource optimization is subject to the number of available resources K, the number of devices M, and the number of eNBs N by adopting joint resource allocation among all cells. Such complexity is around the level of $\mathcal{O}(KMN)$. However, the information-bridled resource optimization only optimizes $\mathbf{R} = [r_1, r_2, \ldots, r_K]$. The complexity is thus $\mathcal{O}(K)$. Please note that information-bridled resource optimization is very different from closed-loop communications in the following aspects.

- In closed-loop communications, the receiver feeds back CSI to the corresponding transmitter. However, the H-CRAN broadcasts a common $\mathbf{R}$ to all transmitters, which does not impose large signaling overheads.
- The concept of information-bridled resource optimization is infeasible to apply to closed-

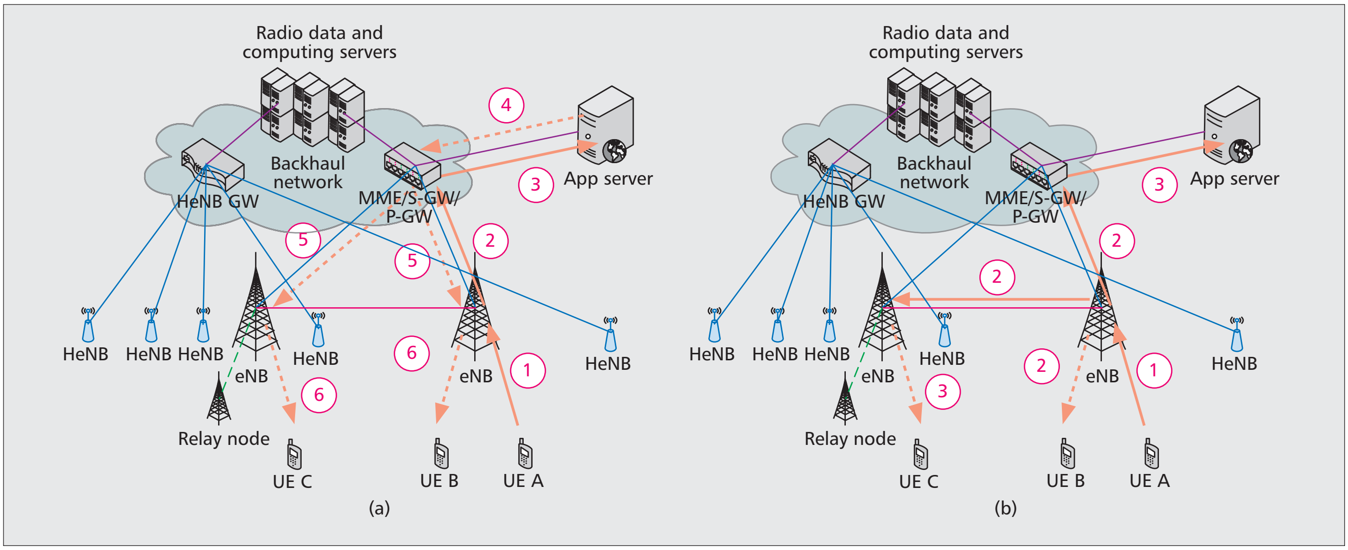


Figure 5. a) In existing mobile networks, six segments are needed to exchange an application message between two mobile devices; b) by caching social profiles of mobile devices in the H-CRAN, the number of hops (and thus latency) can be significantly reduced.

loop communications. In closed-loop communications, the optimization scheme in layers 1 and 2 is fixed to maximize the data rate based on the present SINR. However, the H-CRAN controls the SINR via radio resource allocations, and thus in this framework, the H-CRAN cannot further reduce the complexity.

- In closed-loop communications, layers 1 and 2 of a transmitter need instantaneous CSI. However, exchanging information within the H-CRAN suffers from inevitable latency, which may not support instantaneous CSI exchanges in closed-loop communications.

Social Data Cache-Based Routing/Paging

To further alleviate the third type of latency in routing and paging in the H-CRAN, we should understand the state-of-the-art procedure in existing mobile networks, shown in Fig. 5a. Most wireless services (e.g., user social applications, M2M IoT communications, and remote control), especially for full automation, need an application (app) server to store wireless service data and handle service functions. Suppose that a mobile device (UE-A) wishes to send a message to (or access) another mobile device within the same eNB's coverage (UE-B) or within other eNB's coverage (UE-C); a six-segment protocol is needed for this procedure.

- **1st segment.** UE-A sends the message to the eNB.
- **2nd segment.** The eNB forwards this message to the S-GW/P-GW.
- **3rd segment.** The S-GW/P-GW routes this message to the APP server.
- **4th segment.** Upon receiving the message, the APP server stores this message, then routes this message back to the S-GW/P-GW.
- **5th segment.** According to paging information, the S-GW/P-GW forwards the message to the eNB.
- **6th segment.** The eNB sends the message to UE-B (or UE-C).

Such a framework may invoke unacceptable latency in the worst case, which is not allowed by the remote control to robots/vehicles, intelligent transportation systems, and immersive sensory experience. Nevertheless, the number of hops in Fig. 5a can actually be significantly reduced, as shown in Fig. 5b.

- **1st segment.** UE-A sends the message to the eNB.
- **2nd segment.** If the H-CRAN caches the knowledge about the destination of the message, the eNB is able to directly forward the message to the collocated mobile device (UE-B) or forward to another eNB with the destination mobile device (UE-C). In the meantime, the eNB also routes the message to the S-GW/P-GW.
- **3rd segment.** The eNB forwards the message to UE-C. The S-GW/P-GW forwards the message to the app server.

The challenge of applying the state-of-the-art framework in Fig. 5a to the enhancement in Fig. 5b is twofold.

•The existing framework assumes that a mobile device may send messages to every mobile device in the world with equal likelihood. For this purpose, the sophisticated routing/paging architecture as shown in Fig. 5a shall be adopted. However, this assumption is impractical due to disregard of the social relationship among mobile devices [11]. The social relationship is the correlation among mobile devices in terms of geographic locations (i.e., collocated with each other), identities (e.g., students of the same university), users' interpersonal connections, or contact platforms (e.g., connect to a common website, email/game server). In other words, the destinations (i.e., mobile devices, servers, and websites) a mobile device contacts are subject to the social network or social properties of the mobile device user. According to different communication categories, there are three types of social networks: human-to-human, human-to-machine, and M2M (i.e., interaction of machines), as shown in Fig. 6a. In all these three types, each mobile device only contacts a certain set of devices instead of contacting all devices in the network system. Such a social profile of each mobile device does not change rapidly, and this nature is not fully exploited in the existing design of mobile networks [12, 13].

•The social profile of each mobile device is available in the application layer. However, this information is unavailable to radio layers based on existing layered architecture of mobile networks.

The general concept of a social network driven mobile network design first arose in [11], which revealed a number of novel methodologies for future mobile networks. However, to develop a solution compatible to the H-CRAN, the impacts on the existing infrastructures should be minimized. To combat this engineering constraint, a promising design lies in the concept of OpenFlow in software-defined networking (SDN) [14, 15]. In OpenFlow, the network is able to extract information in packet headers to distinguish data packets and control packets with different QoS requirements, and impose information to packet headers to boost packet routing. However, to enable the performance enhancement in Fig. 5b, a co-design between SDN in the H-CRAN and the app server is needed.

•The H-CRAN caches the social profiles of all mobile devices provided by app servers, as shown in Fig. 6a. Based on these social profiles, the H-CRAN can construct a connection map for each mobile device in OpenFlow. The connection map specifies the route to possible destinations of a mobile device. Since the number of possible destinations for a mobile device is limited to the cardinality of the mobile device's social network, the size of a connection map is also limited. The construction of a connection map also takes into account the radio resource allocation in the H-CRAN. For example, if two mobile devices are collocated, a D2D link can be configured by the H-CRAN for direct data exchanges among mobile devices.

•When a mobile device sends a message, the app server allows the H-CRAN to directly forward the message to the destination without passing through the app server. Nevertheless, the H-CRAN could forward the message to the app server whenever necessary.

The performance of the above social data cache-based routing/paging scheme can be primitively evaluated via the routing process using mobile IP. In this experiment, a router (home network) is connected by two devices (say, DEV1 and DEV2). Another router (foreign network) is connected by the home network router and a device (DEV3), as shown in Fig. 6b. In the conventional routing procedure, when DEV1 wishes to communicate with DEV2, the home network router checks whether DEV2 is within its routing domain. If it is true, packets from DEV1 are forwarded to DEV2. However, in the social data cache-based routing scheme, the social profile of DEV1 (in this case DEV2) is available for the home network router. As DEV1 attaches to the home network router, a routing table to DEV2 is ready in the home network router, and therefore the routing latency is significantly reduced. Such performance is demonstrated in Fig. 6c, where routing latency is shown in time units. A time unit is a logical unit of time used by counters in routing protocols. If DEV1 wishes to communicate with DEV3, in the conventional routing procedure, the home network router checks whether DEV3 is within its routing domain. If it is not true, the home network router checks whether DEV3 is within the foreign network router's routing domain. If it is true, packets of DEV1 can be routed via the home network router and the foreign network router to DEV3. On the contrary, in the social data cache-based routing scheme, since both the home network router and foreign network router have knowledge of the social relationship between DEV1 and DEV3 (i.e., social profile), when DEV1 attaches to the home network router, a routing table to DEV3 is ready for DEV1. This technical merit of the social data cache-based routing scheme can be observed from Fig. 6c.

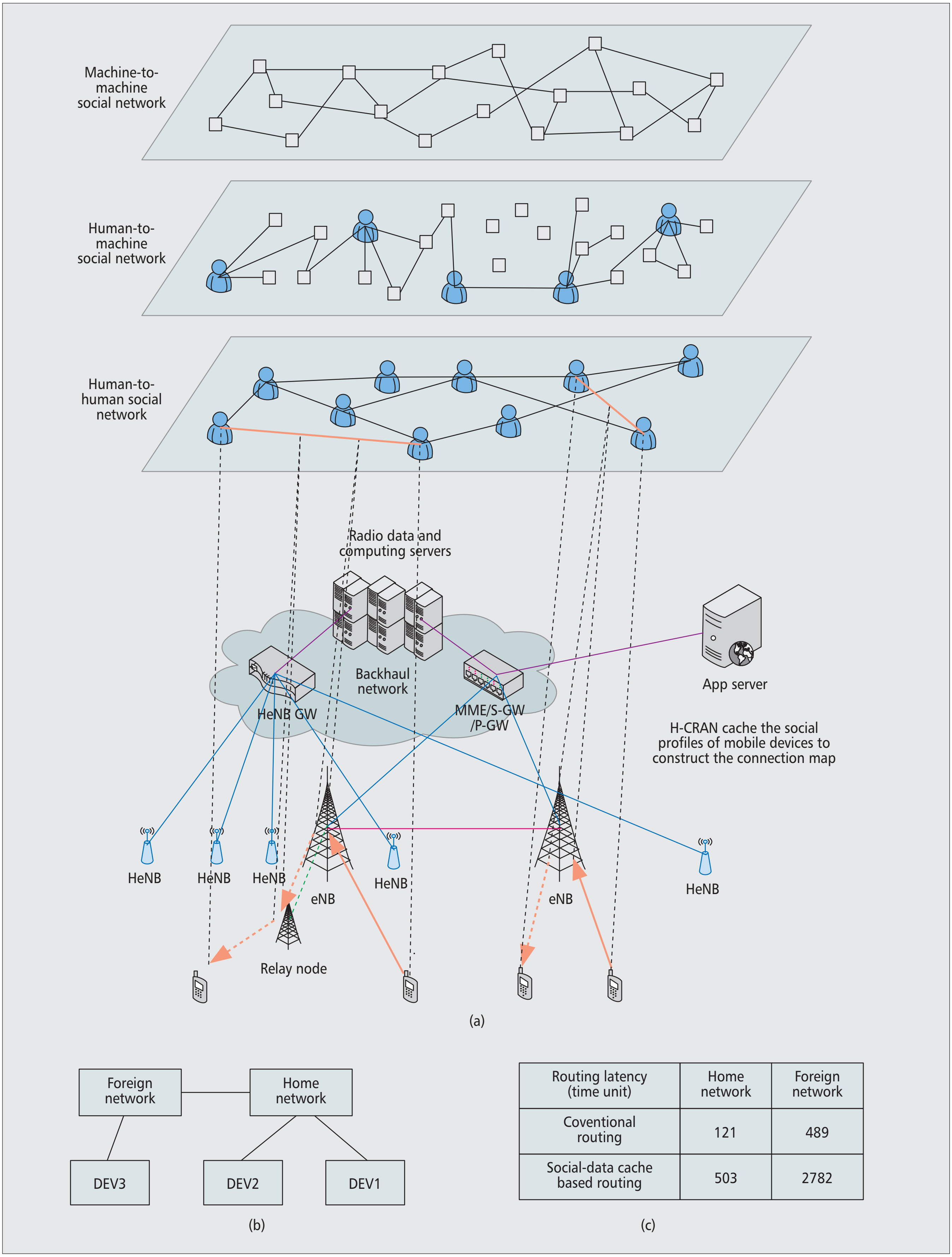
The concept of the above social data cache-based routing/paging scheme significantly reduces packet exchanges crossing the H-CRAN and the cyber world. It also precludes destinations outside the social network of a mobile device, which leads to a concise routing/paging eliminating latency in the H-CRAN.

The concept of the above social data cache-based routing/paging scheme significantly reduces packet exchanges crossing the H-CRAN and the cyber-world. It also precludes destinations outside the social network of a mobile device, which leads to concise routing/paging, eliminating latency in the H-CRAN.

Conclusion and Future Research

In this article, we have presented methodologies enabling ultra-low-latency connections in the H-CRAN, which involve a systematic design with unique open-loop radio access reducing latency in the air interface, information-bridled resource optimization reducing latency of radio resource optimization, and social data cache-based routing/paging scheme reducing latency in the backhaul packet forwarding. The proposed methodologies solve several predicaments and open issues in the H-CRAN through introducing a new design philosophy into the practical H-CRAN.

The future extension of this research is to support *heterogeneous carrier communications* over the H-CRAN. The urgent need for heterogeneous carrier communications comes from the increasing demand for wider bandwidth, which drives the exploration of spectrum with very different characteristics from those of the carrier currently in use. Promising paradigms of heterogeneous carrier communications lie in licensed-assisted access (LAA) to the unlicensed bands in 3GPP Rel-13, millimeter-wave (mmWave) communications, and IEEE 802.11af. In LAA, in addition to conventional licensed bands, 5 GHz unlicensed bands are further included for data transmissions. However, due to the limits on the maximum transmission power and uncontrollable interference from WiFi, communications over the unlicensed bands are potentially unreliable. As a result, carrier aggregation with the control channels provided by the macrocell eNB on the licensed bands and data channels provided by the femtocell HeNB on the unlicensed bands is the mandatory function for LAA. mmWave communications with severe signal


Machine-to-machine social network
Human-to-machine social network
Human-to-human social network
Radio data and computing servers
Backhaul network
HeNB GW
MME/S-GW /P-GW
App server
H-CRAN cache the social profiles of mobile devices to construct the connection map
HeNB
HeNB
HeNB
HeNB
HeNB
eNB
eNB
Relay node
(a)
Foreign network
Home network
DEV3
DEV2
DEV1
(b)
(c)

Routing latency (time unit)	Home network	Foreign network
Coventional routing	121	489
Social-data cache based routing	503	2782

Figure 6. a) The H-CRAN caches social profiles to construct connection maps for all mobile devices; b) system layout for the performance of the proposed social data cache-based routing/paging scheme; c) performance evaluation results of the conventional scheme and the proposed scheme.

strength attenuation on the high frequency bands and IEEE 802.11af operating on the TV bands also suffer from the similar obstacle of communications unreliability. Coordination among cells (HeNBs, eNBs, or access points) and thus the H-CRAN architecture therefore turn out to be the underlay for heterogeneous carrier communications. To support H-CRAN empowered heterogeneous carrier communications, the most challenging issue of communication unreliability may significantly impact the latency performance. Our future research will consequently focus on analyzing and solving this new approach to latency enhancement.

To support H-CRAN empowered heterogeneous carrier communications, the most challenging issue of communication unreliability may significantly impact the latency performance. Our future research will consequently focus on analyzing and solving this new approach to latency enhancement.

Acknowledgment

This research is supported by the Ministry of Science and Technology and MediaTek under contract MOST 103-2622-E-002-034-.

References

[1] D. Lopes-Perez *et al.*, "Enhanced Intercell Interference Coordination Challenges in Heterogeneous Networks," *IEEE Commun. Mag.*, vol. 18, no. 3, June 2011, pp. 22–30.

[2] M. Peng *et al.*, "Heterogeneous Cloud Radio Access Networks: A New Perspective for Enhancing Spectral and Energy Efficiencies," *IEEE Wireless Commun.*, vol. 12, no. 6, Dec. 2014, pp. 126–35.

[3] M. Peng *et al.*, "Energy-Efficient Resource Assignment and Power Allocation in Heterogeneous Cloud Radio Access Networks," to appear, *IEEE Trans. Vehic. Tech*.

[4] L. Lei *et al.*, "Challenges on Wireless Heterogeneous Networks for Mobile Cloud Computing," *IEEE Wireless Commun.*, vol. 20, no. 3, June 2013, pp. 34–44.

[5] R. Balakrishnan and B. Canberk, "Traffic-Aware QoS Provisioning and Admission Control in OFDMA Hybrid Small Cells," *IEEE Trans. Vehic. Tech.*, vol. 63, no. 2, Feb. 2014, pp. 802–10.

[6] D. Chen, T. Q. S. Quek, and M. Kountouris, "Backhauling in Heterogeneous Cellular Networks: Modeling and Tradeoffs," to appear, *IEEE Trans. Wireless Commun.*

[7] I.-W. Lai *et al.*, "End-to-End Virtual MIMO Transmission in Ad Hoc Cognitive Radio Networks," *IEEE Trans. Wireless Commun.*, vol. 13, no. 1, Jan. 2014, pp. 330–41.

[8] Y.-C. Liang *et al.*, "Cognitive Radio Networking and Communications: An Overview," *IEEE Trans. Vehic. Tech.*, vol. 60, no. 7, Sept. 2011, pp. 3386–3407.

[9] S.-Y. Lien *et al.*, "Cognitive radio Resource Management for Future Cellular Networks," *IEEE Wireless Commun.*, vol. 21, no. 1, pp. 70–79, Feb. 2014.

[10] S.-Y. Lien *et al.*, "Radio Resource Management for QoS Guarantees in Cyber-Physical Systems," *IEEE Trans. Parallel Distrib. Sys.*, vol. 23, no. 9, Sep. 2012, pp. 1752–61.

[11] K.-C. Chen, M. Chiang, and H. V. Poor, "From Technological Networks to Social Networks," *IEEE JSAC*, vol. 31, no. 8, Aug. 2013, pp. 1–26.

[12] Y. Yang and T. Q. S. Quek, "Optimal Subsides for Shared Small Cell Networks — A Social Network Perspective," *IEEE J. Sel. Topics Signal Processing*, vol. 8, no. 4, Aug. 2014, pp. 690–702.

[13] E. Stai, V. Karyotis, and S. Papavassiliou, "Exploiting Social-Physical Network Interactions via A Utility-Based Framework for Resource Management in Mobile Social Networks," *IEEE Wireless Commun.*, vol. 21, no. 1, Feb. 2014, pp. 10–17.

[14] H. Kim and N. Feamster, "Improving Network Management with Software Defined Networking," *IEEE Commun. Mag.*, vol. 51, no. 2, Feb. 2013, pp. 114–19.

[15] S. H. Yeganeh, A. Tootoonchian, and Y. Ganjali, "On Scalability of Software-Defined Networking," *IEEE Commun. Mag.*, vol. 51, no. 2, Feb. 2013, pp. 136–41.

Biographies

Shao-Yu Lien (sylien@nfu.edu.tw) is an assistant professor at the Department of Electronic Engineering, National Formosa University, Taiwan. He as received a number of prestigious recognitions, including the IEEE Communications Society Asia-Pacific Outstanding Paper Award in 2014, Scopus Young Researcher Award (issued by Elsevier) in 2014, URSI AP-RASC 2013 Young Scientist Award, and IEEE ICC 2010 Best Paper Award. His research interests include optimization techniques for networks and communication systems. Recently, focuses are particularly on cyber-physical systems and 5G communication networks.

Shao-Chou Hung (d02942008@ntu.edu.tw) received his B.S. and M.S. degree in electrical engineering from National Taiwan University in 2010 and 2013, respectively. He is currently pursuing his Ph.D. degree at the Graduate Institute of Communication Engineering of National Taiwan University. His research interests include 5G network architecture, cognitive radio networks, and machine learning for multi-agent wireless networks.

Kwang-Cheng Chen [F] (chenkc@cc.ee.ntu.edu.tw) is the Distinguished Professor and Associate Dean for Academic Affairs, College of Electrical Engineering and Computer Science, National Taiwan University. He received the 2011 IEEE ComSoc WTC Recognition Award, 2014 IEEE Jack Neubauer Memorial Award, and 2014 IEEE ComSoc AP Outstanding Paper Award. His research interests include wireless communications, network science, and data analytics.

Ying-Chang Liang (liangyc@ieee.org) is a Principal Scientist at the Institute for Infocomm Research, A*STAR, Singapore. He was a visiting scholar at the Department of Electrical Engineering, Stanford University, California, from December 2002 to December 2003, and was an adjunct staff member with the National University of Singapore and Nanyang Technological University from 2004 to 2009. He was recognized by Thomson Reuters as a Highly Cited Researcher in June 2014. He served as Editor-in-Chief of the *IEEE Journal on Selected Areas in Communications* Cognitive Radio Series, and was the key founder of the new journal *IEEE Transactions on Cognitive Communications and Networking*.