Resource-Optimal Licensed-Assisted Access in Heterogeneous Cloud Radio Access Networks With Heterogeneous Carrier Communications

Shao-Yu Lien, Shin-Ming Cheng, *Member, IEEE*, Kwang-Cheng Chen, *Fellow, IEEE*, and Dong In Kim, *Senior Member, IEEE*

***Abstract*—To support a tremendous amount of traffic demands via wireless access in 2020 and beyond, limited bandwidth of the licensed bands has been a major obstacle to boosting further the capacity of wireless services. To fundamentally break through this predicament, an emerging technology known as heterogeneous carrier communications has been launched into standardization in the form of licensed-assisted access (LAA) to the unlicensed bands. Integrating the heterogeneous cloud radio access networks and carrier aggregation, and although LAA-empowered cellular networks gain wider bandwidth from the unlicensed bands, communications may suffer from intersystem interference. To avoid interference, listen-before-talk has been designated as a mandatory function; however, it leads to significant challenges of the hidden-terminal problem. To address this open issue in LAA, in this paper, we consequently propose a resource-optimal scheme using a minimum amount of replicated radio resources to achieve the most essential latency guarantees for real-time applications. To further support non-real-time applications, a new resource control as well as the mathematical architecture inspired by so-called *political communications* is proposed to further maximize the throughput of packet delivery. Our scheme not only optimizes resource utilization in time and spatial domains but also suggests optimum energy efficiency and computation efficiency, to successfully deploy cellular networks on the unlicensed bands.**

***Index Terms*—Heterogeneous carrier communications, heterogeneous cloud radio access networks (CRANs), licensed-assisted access (LAA), political communications.**

Manuscript received April 17, 2014; revised February 28, 2015 and September 4, 2015; accepted January 12, 2016. Date of publication January 29, 2016; date of current version December 14, 2016. The work of S.-Y. Lien was supported by the Institute for Information Industry, Taiwan, and the MOST Project 104-3115-E-009-004. The work of S.-M. Cheng was supported by the MOST Project 104-2221-E-011-051 and by the Industrial Technology Research Institute, Taiwan. The work of D. I. Kim was supported by the National Research Foundation of Korea funded by the Korean government (Ministry of Science, ICT and Future Planning) under Grant 2013R1A2A2A01067195. The review of this paper was coordinated by Prof. O. B. Akan.

S.-Y. Lien is with the Department of Electronic Engineering, National Formosa University, Yunlin 63201, Taiwan (e-mail: sylien@nfu.edu.tw).

S.-M. Cheng is with the Department of Computer Science and Information Engineering, National Taiwan University of Science and Technology, Taipei 10617, Taiwan (e-mail: smcheng@mail.ntust.edu.tw).

K.-C. Chen is with the Department of Electrical Engineering, Graduate Institute of Communication Engineering, National Taiwan University, Taipei 10617, Taiwan (e-mail: ckc@ntu.edu.tw).

D. I. Kim is with the School of Information and Communication Engineering, Sungkyunkwan University, Suwon 440-746, South Korea (e-mail: dikim@skku.ac.kr).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TVT.2016.2523562

I. Introduction

IT is projected that in 2020 and beyond, there will be hundred billion heterogeneous devices, such as user equipment (UE) devices, sensors, machines, or actuators, relying on cellular networks to exchange data [1]–[3]. Such a tremendous increase in the number of devices and the use of diverse wireless applications subsequently result in an explosive growth in traffic demands via wireless access services [4], [5]. As a consequence, not only does traffic volume overwhelm the capacity of existing cellular networks but also the required high-data-rate communications drain the limited radio resources. In principle, the straightforward solutions to this issue are to broaden the bandwidth of communications. However, the current spectrum allocations leave very limited available bandwidth for extra licensed bands, which decreases the approachability of applying existing solutions to the next-generation cellular networks. Nevertheless, this challenge ascends the interests in developing the technology of heterogeneous carrier communications.

The concept of the heterogeneous carrier communications comes from the idea of deploying cellular networks over the unlicensed bands [6]. This idea extends the bandwidth of cellular networks from the limited licensed bands to the broadband unlicensed spectrum, and the corresponding standardization progress known as licensed-assisted access (LAA) has been launched in Third-Generation Partnership Project Rel-13 since 2014 [7]–[10]. However, this deployment encounters two engineering concerns as follows. 1) On the 5-GHz unlicensed bands, there are two types of wireless systems: IEEE 802.11a/ac/ax (Wi-Fi) and weather radars. These existing systems may invoke interference to LAA, and such interference cannot be controlled by LAA. As a result, communications on the unlicensed bands may be unreliable. 2) There is an upper limit of transmission power on the unlicensed bands. This regulation fundamentally shortens the communication range and limits the "cell size" on the unlicensed bands, together with uncoordinated interference [11]. A small cell size is very unfavorable for mobility management in cellular networks, since it severely increases the number of handovers for mobile devices. To eliminate these two concerns, a special design of LAA is adopted by Rel-13 to allocate the control channels to the conventional licensed bands, whereas the data channels are allocated to the unlicensed bands. The merit of this design is twofold. First, since communications on the unlicensed bands may suffer interference from

other wireless systems, exchanging control signalings on the licensed bands is more reliable to facilitate network and resource management. Second, since larger transmission power can be applied to the licensed bands, the communication range is thus longer than that on the unlicensed bands. Allocating control channels on the licensed bands thus decreases the number of handovers. This design well explains the name of "LAA" in Rel-13, which further reveals that an efficient system architecture supporting the heterogeneous carrier communications lies in the heterogeneous cloud radio access networks (CRANs).

In the conventional evolved NodeB (eNB) operation, radio heads and baseband units (BBUs) are embedded (collocated) in an eNB. Under this framework, although the control and data channels are allocated to the licensed and unlicensed bands, respectively, the distances among eNBs for the cell planning are still required to be very limited to provide the coverage continuity of the data channels for LAA. This framework does not alleviate the number of handovers in LAA. However, in the CRAN, multiple remote radio heads (RRHs) and BBUs can be deployed far apart from an eNB [12]–[15]. Via fiber-optical cables to connect all RRUs/BBUs to an eNB, radio resources of each RRH/BBU are scheduled and allocated by the eNB using the cloud computing technology [16], [17]. The CRANs thus allow a dense deployment of RRHs/BBUs, while maintaining a large distance among eNBs, as shown in Fig. 1(a). With this technical merit, radio resources for data exchanges (data channels) at different RRHs/BBUs can be regarded as a unified radio resource pool. Thus, different RRHs/BBUs may be transparent to UE devices and can be viewed as a single "big cell." By operating the control and data channels on eNBs and RRHs/BBUs, respectively, the number of handovers in LAA can thus be effectively suppressed. Furthermore, the capability of RRHs/BBUs has been largely enhanced recently due to the technological maturity of heterogeneous networks [18]–[23] composed of macrocells, femtocells, picocells, and relay nodes (RNs). By connecting home eNBs (HeNBs), RNs, or UE devices (with relay capability) to eNBs via wired (S1 and X2 interfaces) or wireless (Un) backhauls, the CRAN can thus be extended to the heterogeneous CRAN [7], [24], as shown in Fig. 1(b).

In spite of the facilitation of the heterogeneous CRAN, an open issue that obstructs the practice of LAA is known as the *LAA-WiFi hidden-terminal problem* [25]–[27]. Since the data channels of LAA are allocated to the unlicensed bands, to alleviate interference to/from other collocated wireless systems, each transmitter (i.e., eNB, RN, or UE) needs to sense the channel, and transmissions can take place only if the channel is sensed to be idle. This mechanism is referred to as listen-before-talk (LBT) in LAA [7], which has been a mandatory function for radio access services on the unlicensed bands in Japan and Europe [28]. However, since an LAA transmitter and an LAA receiver may be geographically separated apart, a clear channel sensed at the LAA transmitter side does not imply that the channel is also clear at the LAA receiver side. Therefore, data transmissions from a transmitter may suffer from severe interference at the receiver side, although the channel is clear at the transmitter side. This phenomenon is thus the *LAA-WiFi hidden-terminal problem*. In Wi-Fi networks, request-to-send (RTS) and clear-to-send (CTS) messages are exploited to

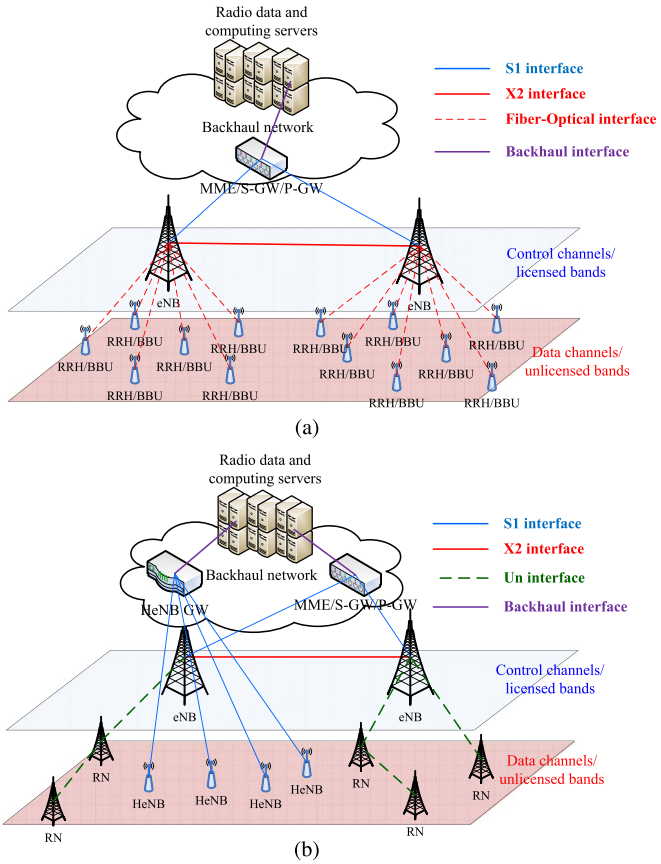

Fig. 1. Effective architecture to deploy LAA lies in the CRAN or the heterogeneous CRAN. (a) In the CRAN, multiple RRHs/BBUs can be densely deployed apart from the eNB. By allocating control channels to the licensed bands through the eNB and allocating data channels to the unlicensed bands through RRHs/BBUs, the number of handovers can be alleviated to support cellular networks. (b) The CRAN can be extended to the heterogeneous CRAN.

alleviate the hidden-terminal problem. However, without a universal air interface for information exchanges between a Wi-Fi network and an LAA network, this RTS-CTS exchange scheme cannot be utilized to avoid the LAA-WiFi hidden-terminal problem, which leads to a challenging consequence. For uplink transmissions in LAA, an eNB schedules radio resources to a UE device to upload data. Nevertheless, a UE device needs to perform LBT on the allocated radio resources and only utilizes the radio resources without interference. As a result, if a UE device requests a certain amount of radio resources and an eNB allocates the exact amount of radio resources requested by the UE, these radio resources may not be fully utilized by a UE device due to interference. If some allocated radio resources suffer from interference, then transmissions on these resources shall be suspended. It substantially harms the latency performance to support real-time applications. To defeat this open issue, an eNB may thus allocate more radio resources than the amount requested by a UE device, and the latency performance could be improved if these replicated radio resources are not all interfered with simultaneously. For this purpose, we note that the heterogeneous CRANs provide a particular technical feature of multipath transmissions, as shown in Fig. 2. Specifically, for each UE device, there can be multiple communication paths to forward packets to the

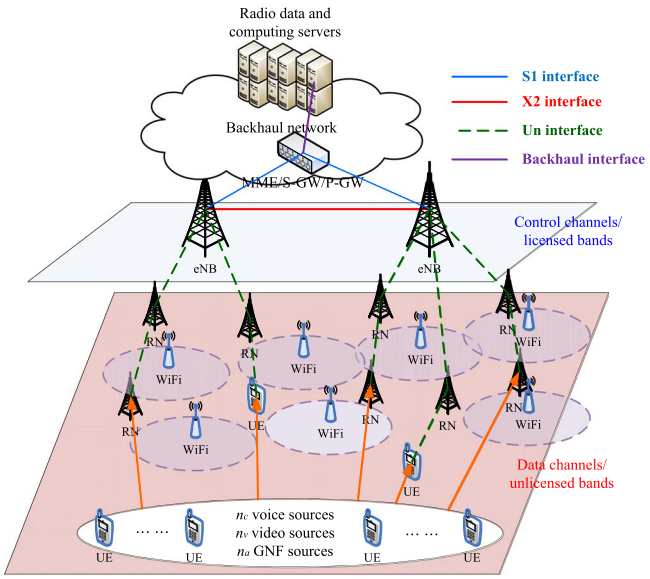

Fig. 2. In the heterogeneous CRAN, there can be K disjoint paths formed by RNs and eNBs, where each path is composed of L_k links suffering different levels of interference.

eNB(s). For real-time applications, if each packet transmission is repeated via multiple paths, then the timing constraint (delay or jitter) of the packet transmission is violated only if all paths suffer from severe interference concurrently. This mechanism is referred to as carrier aggregation (CA) [7] in the spatial domain in LAA, which thus significantly alleviates the probability of timing constraint violation. However, this CA operation reveals a tradeoff between the number of utilized paths and the timing constraint violation probability. Considering that the transmission repetition leads to significant overheads, to minimize the amount of allocated radio resources, we shall thus strike the optimum tradeoff in this paper.

To support packet transmissions of the general non-real-time file (GNF), the inherent LBT results in a new challenge different from that for supporting real-time applications. Since there is no timing constraint for packet transmissions of the GNF, it is not necessary to transmit replicates of each GNF packet via multiple paths. Instead, forwarding each GNF packet via only one communication path turns out to be an efficient scheme. Given that k_a communication paths are required to support real-time applications, a total of k_a GNF packets can be simultaneously delivered via k_a paths to fully utilize all communication paths. However, since LBT shall be performed at each path, if a path is allocated to a particular UE device, transmissions cannot take place at this path if the channel is sensed busy at this path. In this case, the throughput of a UE device is degraded. To further enhance the throughput of a UE device, an eNB may allocate only one path to a particular UE device. Instead, k_a communication paths are regarded as a resource pool [10]. When a UE device attempts to upload data, this UE performs LBT at all communication paths and selects one path with the channel sensed to be idle. If each path suffers from different levels of interference and all UE devices select different paths, the overall throughput is improved. However, the worst case is severe congestion of GNF packets on certain communication paths (i.e., some UE devices select the same path, leading to traffic load imbalance among communication paths). To achieve optimum load balance (thus, optimum resource utilization), an innovative design is required to yield optimization with extremely low complexity. For this purpose, an interdisciplinary principle of "political communications" [29] is adopted to trace the throughput of all GNF packet transmissions to the optimum. The spirit of political communications originates from the optimization theory that the performance of an optimization is subject to available information, and thus, the performance of an optimization can be controlled by bridling available information. By exploiting this concept, if an eNB reveals its congestion levels at each path to all UE devices, then each UE is able to select a proper path without severe congestion. If an eNB further optimizes the revealed congestion levels, then each UE is able to further optimize the individual path selection to maximize the overall throughput.

To successfully practice heterogeneous carrier communications in the form of LAA, in this paper, we shall solve the aforementioned open issues in LAA. Our resource-optimal scheme first utilizes the minimum amount of paths to provide latency guarantees for real-time applications. Then, given the minimum amount of paths, we develop a mathematical framework to optimize the load balance of all communications paths via optimizing the available information announced to all UE devices. Our scheme consequently imposes practicable complexity, which not only achieves optimum resource utilization in the time domain and in the spatial (path) domain but also suggests optimized energy efficiency (as a minimum number of paths is involved).

II. System Model

In this paper, we consider the uplink data transmissions in LAA, where the data channels are allocated on the unlicensed bands. With the facilitation of the heterogeneous CRAN, eNBs and RNs are able to cooperate with each other to form K disjoint communication paths, which are indexed by $k = 1, \ldots, K$. Each of these K paths (e.g., the kth path) is composed of L_k links. As each of these L_k links may be shared by other wireless systems, they may suffer from different levels of interference, as shown in Fig. 2. For uplink transmissions, eNBs are receivers of each path, whereas UE devices are transmitters of each path. At each link, RNs or UE devices for data relays first receive data (receivers) from the previous link and then transmit data (transmitters) to the subsequent link. At each link, each transmitter needs to perform LBT.

A. Consideration of LBT

In Europe, two kinds of LBT schemes are defined [28].

- **Load-Based Equipment (LBE)**: Before a transmission on the channel, the equipment shall perform a *clear channel assessment (CCA)* check using "energy detection." The equipment shall observe the channel for a duration of the CCA observation time (> 20 μs). The channel

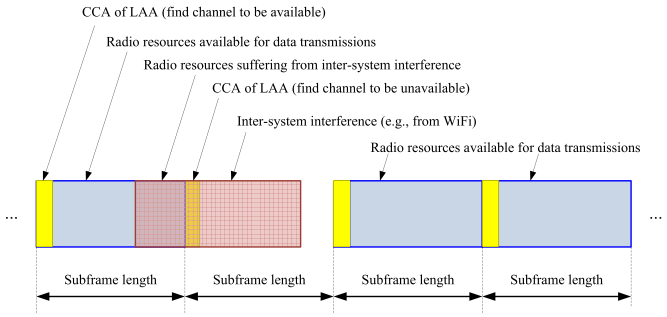

Fig. 3. By applying LAA for communications on the unlicensed bands, some radio resources may be occupied by other systems as interference. To mitigate interference, the heterogeneous CRAN should adopt LBT (specifically, FBE) and link adaptation schemes. As a result, the number of available radio resources is dynamic.

is considered to be occupied if the energy level in the channel exceeds a certain threshold. If the equipment finds the channel to be clear, it may transmit immediately. However, if the equipment finds the channel to be occupied, it shall not transmit on the channel. The equipment shall perform a CCA check for a duration of a random factor Y multiplied by the CCA observation time. Y is stored in a counter, which is decremented every time a CCA observation time is regarded to be "unoccupied." When the counter reaches zero, the equipment may transmit. When an equipment device successfully occupies the channel, the total time that the equipment can transmit on the channel is a *maximum channel occupation time*.

- **Frame-Based Equipment (FBE)**: The operation of FBE is basically the same as that of LBE, but there is one difference. In FBE, if the equipment finds the channel to be occupied, the equipment waits for a fixed period of time (referred to as a channel occupation time) and performs a CCA check again. If the equipment finds a clear channel, then it can transmit on the channel for a channel occupation time.

In this paper, FBE is adopted due to its technical merit of simplicity. By adopting FBE, each UE and RN shall perform CCA at every subframe for at least 20 μs. If the channel is sensed to be clear, a UE device or an RN is able to transmit in a subframe. Otherwise, a UE device or an RN shall suspend the transmission in a subframe, as shown in Fig. 3. By adopting FBE, the single-carrier frequency-division multiple access adopted by Long-Term Evolution Advanced (LTE-A) for uplink transmissions thus degenerates to a time-division multiplexing operation. To maintain a prescribed packet error rate (PER), the link adaptation schemes are typically adopted to maximize the transmission rates based on the present signal-to-interference-plus-noise power ratio on the link [30]–[32]. As a result, after performing LBT at the transmitter side of each link, there can be $Z+1$ transmission modes (transmission rates) at the transmitter side of each link, and the probability distribution of a particular transmission mode among $Z+1$ on the lth link of the kth path can be modeled by

$$\Pi_{l,k} = \left[\pi_0^{l,k}, \pi_1^{l,k}, \ldots, \pi_Z^{l,k}\right], \ \sum_{z=0}^{Z} \pi_z^{l,k} = 1 \tag{1}$$

where $\pi_0^{l,k}$ is the probability that the link is regarded to be occupied, and

$$\varphi_{l,k} = \sum_{z=1}^{Z} \pi_z^{l,k} \tag{2}$$

is the probability that the lth link of the kth path at each subframe is regarded not to be occupied. Therefore, it may need multiple subframes to deliver a packet through the lth link of the kth path. However, due to the LAA-WiFi hidden-terminal problem, when a transmission mode is applied to the transmitter side of a link as the channel is sensed to be available, the receiver could suffer from unacceptable interference (i.e., the PER is unacceptable). Denote the probability of unacceptable interference $\pi_{\mathrm{rx}}^{l,k}$ at the receiver side of a link during the subframes to deliver a packet through a link. It takes $\lceil \chi / R_{l,k} \rceil$ subframes to deliver a packet through a link, if there is no interference at the receiver side with the probability $1-\pi_{\mathrm{rx}}^{l,k}$, where χ is the packet size, and $R_{l,k}$ (bits/subframe) is the transmission rate on the lth link of the kth path. On the other hand, if unacceptable interference occurs at the receiver side of a link with the probability $\pi_{\mathrm{rx}}^{l,k}$, then the packet cannot be successfully delivered through this link (and this path). We will elaborate later in Section III that τ_c subframes are allocated for each packet to be transmitted from a source to the final destination via multiple paths with multiple links. In the case with interference at the receiver side of a link, it thus takes τ_c subframes to identify a packet reception failure. The number of subframes to deliver a packet through a link, i.e., $S_{l,k}$, is therefore a random variable with the distribution, i.e.,

$$\Pr\{S_{l,k}\} = \begin{cases} 1-\pi_{\mathrm{rx}}^{l,k}, & S_{l,k} = \left\lceil \frac{\chi}{R_{l,k}} \right\rceil \\ \pi_{\mathrm{rx}}^{l,k}, & S_{l,k} = \tau_c. \end{cases} \tag{3}$$

Thus, $\varphi_{l,k}(1-\pi_{\mathrm{rx}}^{l,k})$ can be regarded as the availability (communication opportunity) on a link, whereas $1-\varphi_{l,k}(1-\pi_{\mathrm{rx}}^{l,k})$ can be regarded as the unavailability on a link. To facilitate the establishment of our scheme, important notations adopted in this paper are summarized in Table I.

B. Packet Sources

In this paper, three types of traffic are considered for LAA: real-time voice, real-time video, and non-real-time GNF. In the literature, it has been shown impossible to provide deterministic latency guarantees (i.e., the probability of the timing constraint violation is zero) over a wireless channel [33]. Consequently, in this paper, statistical performance guarantees for voice and video sources are considered (i.e., the probability of the timing constraint violation is upper bounded by a required value).

S1) A voice source is characterized by three parameters $(\lambda, \delta, \varepsilon)$, where λ is the packet arrival rate of the source, δ is the maximum tolerable jitter, and ε is the acceptable jitter constraint violation probability. Packets of a voice source are periodically generated every $1/\lambda$ subframes and are stored in a ready-to-transmit (RTT) buffer for this voice source. Jitter is defined as the difference

TABLE I
IMPORTANT NOTATIONS

Notation	Definition
K	Total number of communication paths
L_k	Number of links in the kth path
$Z+1$	Number of transmission modes in Layer 1
$\pi_z^{l,k}$	Probability of the zth transmission mode on the lth link in the kth path
$\varphi_{l,k}$	Availability (available probability) of the lth link in the kth path
χ	Size of a packet
$R_{l,k}$	Transmission rate of the lth link in the kth path
$S_{i,k}$	Number of subframes to deliver a packet through the lth link in the kth path
λ_i	Packet arrival rate of the ith voice source
δ_i	Maximum tolerable jitter of the ith voice source
ε_i	Acceptable jitter constraint violation probability of the ith voice source
ρ_j	Average packet arrival of the jth video source
σ_j	Maximum burstness of the jth video source
d_j	Maximum tolerable delay of the jth video source
ξ_j	Acceptable delay constraint violation probability of the jth video source
k_a	Number of communication paths adopted for multipath transmissions
n_c, n_v, n_a	Numbers of voice, video, and GNF sources
$q_1, \ldots, q_{k_a}$	Barrier parameters for each of k_a communication paths
Θ_c^k	True end-to-end packet forwarding time of a voice packet via the kth path
Θ_c	True end-to-end packet forwarding time of a voice packet by leveraging k_a paths
Θ_v^k	True end-to-end packet forwarding time of a video packet via the kth path
Θ_v	True end-to-end packet forwarding time of a video packet by leveraging k_a paths
τ_c	Number of subframes reserved for one packet transmissions
$\bar{\Theta}_c$	Expected value of Θ_c
$\bar{\Theta}_v$	Expected value of Θ_v
$p_{k,f}$	$\Pr\{\Theta_c^k = f\}$
Φ_k	$\Pr\{\Theta_c^k > \tau_c\}$
$S_{l,k}$	Number of subframes to deliver a voice packet through the lth link of the kth path
N_k	Set of GNF sources selecting the kth path
$\|N_k\|$	Number of elements in N_k
$\beta_n(q_1, q_2, \ldots, q_{k_a^*})$	Path selection strategy of the nth GNF source when $q_1, q_2, \ldots, q_{k_a^*}$ are received
$\|N_k^{(t)}\|$	Number of GNF sources selecting the kth paths after the tth iterations in Algorithm 1
$\pi_{rx}^{l,k}$	Probability of unacceptable interference at the receiver side of a link

between the time of two successive packet departures and the time of two successive packet arrivals. Voice sources with a higher arrival rate, i.e., λ, have a higher priority.

S2) A video source is characterized by four parameters (ρ, σ, d, ξ), where ρ is the average packet arrival rate of the source, σ is the maximum burstness (the maximum number of packets in one arrival), d is the maximum tolerable delay, and ξ is the acceptable delay constraint violation probability. A video source regulated by a (σ, ρ)-leaky bucket is stored in an RTT buffer for this video source. Video sources are with bulk arrivals (that is, multiple packets from upper layers may arrive at the same time). Data are decodable at the destination only when the entire bulk of packets is successfully received before the expiration of d. Video sources with a smaller d have a higher priority.

S3) A GNF source does not have a timing constraint. All GNF sources fairly and efficiently share the remaining resources from voice and video sources.

In practical scenarios, each device can be mobile and is able to roam in/out the heterogeneous CRAN. As a result, the number of devices in the heterogeneous CRANs can be fully dynamic. When a device roams in, it can receive communication services from k_a communication paths, where $k_a \leq K$. Each device has a certain number of voice, video, or GNF sources to upload voice, video, and GNF packets, respectively, to the heterogeneous CRAN. By this consideration, without loss of generality, we can avoid the index of individual UE devices to consider a total of n_c voice sources indexed by $i = 1, \ldots, n_c$, n_v video sources indexed by $j = 1, \ldots, n_v$, and n_a GNF sources in the heterogeneous CRAN.

C. Preliminary of the Resource-Optimal Scheme

Although the heterogeneous CRANs support multiple communication paths against communication unreliability due to interference, it also involves multiple transmitters and multiple links. This framework thus heavily consumes not only resources in the time and spatial domains but also energy in UE devices, RNs, and eNBs. To optimize resource utilization, we shall minimize the number of involved communication paths. To achieve this goal, while providing latency guarantees for voice and video sources, an essential requirement is to fully utilize resources in the time domain. Without considering link unavailability and multiple communication paths, it has been shown that the optimum arrangement of voice, video, and GNF sources in the time domain satisfies the following operation principle [34].

1) At the end of a packet-forwarding period, an active voice source (that is, the source has a packet in the RTT buffer) with the highest present priority is allowed to transmit, at most, one packet from its RTT buffer.

2) At the end of a packet-forwarding period, if there is no active voice source, an active video source with the highest present priority is allowed to transmit all packets in its RTT buffer.
3) At the end of a packet-forwarding period, if there is no active voice and video source, the remaining time-domain resources are shared by active GNF sources.

The aforementioned operation principle is able to serve as a foundation for the transmission arrangement of voice, video, and GNF sources. However, the aforementioned principle is still insufficient to provide the optimum resource utilization both in the time domain and in the spatial domain on the unlicensed bands suffering from link unavailability. We will discuss this part in the following section.

III. Resource-Optimal Scheme for Voice and Video Sources

To combat interference that may lead to timing constraint violations for voice and video sources, an effective scheme is to forward replicates of each voice and video packet via multiple communication paths simultaneously [35]–[38]. Consequently, the probability of timing constraint violation can be effectively decreased as the number of utilized communication paths (and thus diversity) increases. On the other hand, for GNF sources, it is not necessary to adopt such diversity for packet forwarding. Instead, different communication paths can carry different GNF packets (and thus the multiplexing scheme), and all GNF packets shall be evenly spread over both the time and the spatial domains so as to maximize the throughput. However, as aforementioned in Section I, performing such optimization at the network side may result in unacceptable complexity. To fundamentally break through this obstacle, each GNF source should autonomously select a proper communication path to forward its packet. However, without coordination among GNF sources, multiple GNF sources may select the same communication path that can cause severe packet congestion. To evenly spread the traffic load of all GNF sources over available paths, political communications provide a key concept that each individual (thus, each GNF source) optimizes its decision based on available information. In the considered scenario, information is the current congestion level at each path. Therefore, the decisions of all individuals can be managed via controlling information given to each individual. The current congestion level at each path can be fully captured by the heterogeneous CRAN. By optimizing provided information to all GNF sources, load balance among paths can be achieved. This design begins by a barrier mechanism, and the aforementioned operation principle is specified as follows.

1) At the end of a packet-forwarding period, an active voice source with the highest present priority is allowed to transmit, at most, one packet from its RTT buffer.
2) At the end of a packet-forwarding period, if there is no active voice source, an active video source with the highest present priority is allowed to transmit all packets in its RTT buffer.
3) If a voice packet or a video packet is forwarded, replicates of this packet are simultaneously forwarded via multiple communication paths.
4) If there is no active voice and active video source, the remaining resources are shared by active GNF sources.
   a) Each GNF packet is forwarded via only one communication path. Therefore, if k_a communication paths are utilized by the heterogeneous CRAN, the maximum of k_a GNF sources can be simultaneously served in a subframe.
   b) The heterogeneous CRANs announce a set of barrier parameters $\mathbf{Q} = [q_1, \ldots, q_{k_a}]$, where $0 \preceq \mathbf{Q} = [q_1, \ldots, q_{k_a}] \preceq 1$, corresponding to each of k_a communication paths. If a particular communication path (e.g., the kth path) is selected by a GNF source to forward its packet, this GNF source is allowed to transmit only its packet with a probability $q' \leq q_k$. As a result, $\mathbf{Q}$ induces GNF sources to spread packets over the time and spatial domains.
5) For all voice, video, and GNF packets, it takes one subframe to forward a packet through a link. If the packet is ready to be forwarded through a link while the link is unavailable, packet forwarding on this link is suspended until the link turns to available.

In the following section, we immediately define the packet-forwarding period in the aforementioned operation principle.

A. Problem Formulation and Analysis

For voice and video sources, although replicates of each packet are via multiple communication paths, the exact number of subframes required to forward a packet from the source to the destination is still unclear. This is because of the unavailability on each link. A practical solution to tackle this issue is to reserve a certain number of subframes for packet delivery of voice and video sources. Specifically, for all voice sources, τ_c subframes are reserved for each packet forwarding. If a bulk of packets from the jth video source is forwarded, $\sigma_j \tau_c$ subframes are reserved for this bulk of packets (as σ_j is the maximum burst of a packet arrival). These reserved subframes are referred to as a *packet-forwarding period*. If a packet-forwarding period is expired while the packets are still not forwarded to the destination, the packets are discarded. However, a dilemma is encountered on the optimization of the packet-forwarding period τ_c. A conservative design may set τ_c to align with the timing constraint of the (voice or video) source. We will particularly evaluate the performance of such design in Section VI. However, if an abundant number of communication paths are utilized, it may only take a few subframes to deliver the packet(s) to the destination. As a result, a conservative design may lead to potential resource wastes in the time domain. On the other hand, if a proactive design is adopted to set τ_c to a small value, it requires a large number of paths to combat link unavailability to provide latency guarantees. This dilemma reveals that the design mentioned in Section III-C is not optimum. For voice and video sources, a resource-optimal scheme for latency guarantees shall optimize

resource utilization both in the time and spatial domains by solving the following optimization.

Definition 1: The true end-to-end packet-forwarding time (in terms of subframes) of a voice packet via the kth communication path, which is denoted by Θ_c^k, is the sum of the number of subframes actually spent for packet forwarding and the number of subframes that the transmission is suspended, to deliver a voice packet from the source to the destination. Consequently, the true end-to-end packet-forwarding time of a voice packet by leveraging k_a paths is

$$\Theta_c = \min\left\{\Theta_c^1, \ldots, \Theta_c^{k_a}\right\}. \tag{4}$$

Definition 2: The true end-to-end packet-forwarding time (in terms of subframes) of a bulk of video packets via the kth communication path, which is denoted by Θ_v^k, is the sum of the number of subframes actually spent for packet forwarding and the number of subframes that transmission is suspended, to deliver a bulk of video packets from the source to the destination. Consequently, the true end-to-end packet-forwarding time of a bulk of video packets by leveraging k_a paths is

$$\Theta_v = \min\left\{\Theta_v^1, \ldots, \Theta_v^{k_a}\right\}. \tag{5}$$

Optimization 1: Denote $(\lambda_i, \delta_i, \varepsilon_i)$ as the parameters of the ith voice source and denote $(\rho_j, \sigma_j, d_j, \xi_j)$ as the parameters of the jth video source. The resource-optimal scheme for voice and video sources is mathematically formulated as

$$\begin{aligned} &\min\ \tau_c k_a \\ &\text{s.t.}\ \ \text{(i)}\ \Pr[\Theta_c > \delta_i] \le \varepsilon_i,\ \text{for } i = 1, \ldots, n_c \\ &\qquad \text{(ii)}\ \Pr[\Theta_v > d_j] \le \xi_j,\ \text{for } j = 1, \ldots, n_v \\ &\qquad \text{(iii)}\ 0 \le k_a \le K. \end{aligned} \tag{6}$$

The objective in (6) minimizes the time-domain resource utilization, i.e., τ_c, and the number of paths k_a, under the timing constraints for voice and video sources in (i) and (ii), and the feasibility constraints of k_a in (iii). However, the feasibility constraint for the selection of τ_c is still unclear. To solve (6), we should analyze the feasibility constraint of τ_c by studying the relationship among τ_c, k_a, timing constraint violation probabilities $\Pr[\Theta_c > \delta_i] \le \varepsilon_i$ for all i, and $\Pr[\Theta_v > d_j] \le \xi_j$ for all j. As voice sources have higher priorities among three classes of sources, this analysis begins from voice sources.

Theorem 1: By utilizing k_a communication paths, denote

$$\delta_i^* = \tau_c + \sum_{g=1}^{i-1} \left\lceil \frac{\lambda_g}{\lambda_i} \right\rceil \tau_c, \quad i = 1, \ldots, n_c \tag{7}$$

where $\lceil x \rceil$ is the integer ceiling of x. If $\delta_i^* + \tau_c \le 1/\lambda_i$ and $\delta_i^* < \delta_i$ for all i, the jitter constraint violation probability of the ith voice source is upper bounded by $\bar{\Theta}_c/\tau_c$, where $\bar{\Theta}_c$ is the expected value of Θ_c.

Proof: Since the packets of the ith voice source are periodically generated every $1/\lambda_i$ subframes, by temporarily assuming $\Theta_c \le \tau_c$, if we can show that the ith voice source has the maximum wait $\tilde{\delta}_i$, the jitter cannot be larger than $\tilde{\delta}_i$. Furthermore, since each packet of voice sources is allocated by τ_c subframes, if $\tilde{\delta}_i + \tau_c < 1/\lambda_i$, the packet can be delivered to the destination before the next packet arrival. We prove the given arguments by induction with two hypotheses, i.e.,

$$\text{i)}\ \tilde{\delta}_i \le \delta_i^* \tag{8}$$

$$\text{ii)}\ \tilde{\delta}_i + \tau_c < 1/\lambda_i. \tag{9}$$

Considering the first voice source, the maximum wait of a packet is $\tilde{\delta}_1 = \tau_c = \delta_1^*$ subframes. To ensure the packet of the first voice source to be delivered to the destination before the next arrival, the sufficient condition is $\tilde{\delta}_1 + \tau_c < 1/\lambda_i$, which is our assumption $\delta_1^* + \tau_c < 1/\lambda_i$. Suppose that the induction hypotheses hold up to the $(i-1)$th voice source. We argue by contradiction that $\tilde{\delta}_i \le \delta_i^*$. Suppose $\tilde{\delta}_i > \delta_i^*$, voice sources $g = 1, \ldots, i-1$ must be served. From induction hypothesis ii), every packet of these $i-1$ voice sources is served before the next packet arrival. Thus, the total number of packets that can be served within $(0, \delta_i^*)$ for these $i-1$ voice sources is, at most, $\sum_{g=1}^{i-1} \lceil \lambda_g \delta_i^* \rceil$. Therefore, the total number of subframes to serve these packets is upper bounded by

$$\sum_{g=1}^{i-1} \lceil \lambda_g \delta_i^* \rceil\, \tau_c + \tau_c. \tag{10}$$

Since $\delta_i^* < 1/\lambda_i$, the quantity in (10) is upper bounded by

$$\sum_{g=1}^{i-1} \left\lceil \frac{\lambda_g}{\lambda_i} \right\rceil \tau_c + \tau_c = \delta_i^* \tag{11}$$

which follows the definition of δ_i^* in (7). Therefore, all k_a paths cannot always be busy in $(0, \delta_i^*)$, and we reach a contradiction. This shows $\tilde{\delta}_i \le \delta_i^*$ and the packets of the ith voice source will be transmitted before the next arrival. These arguments are valid under the assumption $\Theta_c \le \tau_c$. If $\Theta_c > \tau_c$, the packet transmission may violate the maximum tolerable jitter constraint. This probability is denoted by $\Pr[\Theta_c > \tau_c]$, which is consequently upper bounded by $\bar{\Theta}_c/\tau_c$. ■

Theorem 1 fully reveals the relationship between τ_c and $\Pr[\Theta_c > \delta_i]$, by utilizing k_a communication paths. Further incorporating video sources, the following theorem is provided. A video source (e.g., the jth video source) is served by utilizing the remaining time-domain resources after serving all voice sources and previous $j-1$ video sources. Therefore, the maximum delay of the jth video source, i.e., d_j^*, is affected by all voice sources and the maximum delays of previous $j-1$ video sources. Consequently, a recursive form is adopted to result in the following theorem.

Theorem 2: By utilizing k_a communication paths, recursively denote

$$d_j^* = \frac{\Theta_v \left(1 + \sum_{g=1}^{j} \sigma_g + \sum_{g=1}^{j-1} \rho_g d_g^*\right) + \tau_c(1 + n_c)}{1 - \tau_c \sum_{i=1}^{n_c} \lambda_i - \sum_{g=1}^{j-1} \rho_g \sigma_g \tau_c} \tag{12}$$

for $j = 1, \ldots, n_v$. If $\tau_c \sum_{i=1}^{n_c} \lambda_i + \sum_{g=1}^{j-1} \rho_g \sigma_g \tau_c < 1$, then the delay constraint violation probability of the jth video source is upper bounded by $\bar{\Theta}_v/\varpi_j$, where $\bar{\Theta}_v$ is the expected value of

Θ_v, and ϖ_j is given by

$$\varpi_j = \frac{d_j\left(1-\tau_c\sum_{i=1}^{n_c}\lambda_i-\sum_{g=1}^{j-1}\rho_g\sigma_g\tau_c\right)-\tau_c(1+n_c)}{1+\sum_{g=1}^{j}\sigma_g+\sum_{g=1}^{j-1}\rho_g d_g^*}. \tag{13}$$

Proof: Let $C_1(t_1,t_2)$ be the number of subframes that can be allocated to the first voice source in an interval $(t_1,t_2]$. From the proof of Theorem 1, the maximum number of packets from the n_c voice source that can be served in an interval $(t_1,t_2]$ is, at most, $\sum_{i=1}^{n_c}\lceil\lambda_i(t_2-t_1)\rceil$. Applying the following inequality:

$$\lceil x\rceil \le x+1 \tag{14}$$

yields the bound $\sum_{i=1}^{n_c}[\lambda_i(t_2-t_1)+1]$. Since our design is nonpreemptive, the number of subframes that can be allocated to the first video source in $(t_1,t_2]$ is, at least, $t_2-t_1-\tau_c\{1+\sum_{i=1}^{n_c}[\lambda_i(t_2-t_1)+1]\}$. Therefore, we have

$$C_1(t_1,t_2)\ge\left[1-\tau_c\sum_{i=1}^{n_c}\lambda_i\right](t_2-t_1)-\tau_c(n_c+1). \tag{15}$$

Note that the number of departures in $(t_1,t_2]$ from a (σ,ρ)-leaky bucket is upper bounded by $\sigma+\lceil\rho(t_2-t_1)\rceil$. Applying (14) yields the upper bound $\sigma+\rho(t_2-t_1)+1$. Let $A_1(t_1,t_2)$ be the amount of work load (number of subframes required for packets that arrive at the RTT buffer) within the interval $(t_1,t_2]$ for the first video source, then

$$A_1(t_1,t_2)\le\Theta_v\left[\sigma_1+\rho_1(t_2-t_1)+1\right]. \tag{16}$$

The delay of an arrival at time t is upper bounded by $\inf\{d'\ge 0: A_1(0,t)-C_1(0,t+d')\le 0\}$. Maximizing over t, we have

$$d_1^*=\sup_t\inf\left\{d'\ge 0: A_1(0,t)-C_1(0,t+d')\le 0\right\}. \tag{17}$$

Applying the upper constraint of $A_1(t_1,t_2)$ and the lower constraint of $C_1(t_1,t_2)$, we obtain

$$d_1^*=\frac{\Theta_v(1+\sigma_1)+\tau_c(1+n_c)}{1-\tau_c\sum_{i=1}^{n_c}\lambda_i}. \tag{18}$$

If $d_1^*>d_1$, the maximum tolerable delay constraint of the first video source is violated. For the first video source, we obtain

$$\begin{aligned}\Pr\left[d_1^*>d_1\right]&=\Pr\left[\frac{\Theta_v(1+\sigma_1)+\tau_c(1+n_c)}{1-\tau_c\sum_{i=1}^{n_c}\lambda_i}>d_1\right]\\&=\Pr[\Theta_v>\varpi_1]<\frac{\bar{\Theta}_v}{\varpi_1}\end{aligned} \tag{19}$$

where ϖ_1 is defined in (13). This completes the argument for the first video source. The argument for the jth video source is essentially the same as that of the first video source. However, the lower constraint is required to be modified since the jth video source utilizes the remaining (time domain) resources from all voice sources and the first $j-1$ video sources. Parallel to the argument of the first video source, the maximum delay of the jth video source is upper bounded by $(\Theta_v(1+\sum_{g=1}^{j}\sigma_g+\sum_{g=1}^{j-1}\rho_g d_g^*)+\tau_c(1+n_c))/(1-\tau_c\sum_{i=1}^{n_c}\lambda_i-\sum_{g=1}^{j-1}\rho_g\tau_c\sigma_g)$. Therefore

$$\Pr\left[d_j^*>d_j\right]=\Pr[\Theta_v>\varpi_j]<\frac{\bar{\Theta}_v}{\varpi_j}. \tag{20}$$

■

In addition to Theorem 1, Theorem 2 further reveals the relationship between τ_c and $\Pr[\Theta_v>d_j]$, given that k_a communication paths are utilized. With the facilitation of Theorems 1 and 2, Optimization 1 can be rewritten by imposing three additional constraints for the feasibility on the selection of τ_c.

Optimization 2 (Reformulation of Optimization 1): The resource-optimal scheme for voice and video sources is mathematically formulated as

$$\begin{aligned}&\min\ \tau_c k_a\\&\text{s.t.}\ \ \text{(i)}\ \Pr[\Theta_c>\delta_i]\le\varepsilon_i,\ \text{for } i=1,\ldots,n_c\\&\qquad\text{(ii)}\ \Pr[\Theta_v>d_j]\le\xi_j,\ \text{for } j=1,\ldots,n_v\\&\qquad\text{(iii)}\ 0\le k_a\le K\\&\qquad\text{(iv)}\ \delta_i^*+\tau_c\le 1/\lambda_i \text{ and } \delta_i^*<\delta_i,\ \text{for } i=1,\ldots,n_c\\&\qquad\text{(v)}\ \tau_c\left(\sum_{i=1}^{n_c}\lambda_i\right)+\sum_{g=1}^{j-1}\rho_g\tau_c\sigma_g<1,\ \text{for all } i \text{ and } j\\&\qquad\text{(vi)}\ \tau_c>0.\end{aligned} \tag{21}$$

B. Resource-Optimal Scheme Design

Since Optimization 2 is not convex, it is not guaranteed that a locally optimal result suggests the globally optimal result. As a consequence, we need to examine all the feasible solutions, which may lead to unacceptable complexity. Nevertheless, with the facilitation of Theorem 1 and Theorem 2, (iv) forms the strictest condition among (iv), (v), and (vi) in (21). From (iv), the number of feasible choices of τ_c does not exceed $\lceil 1/\lambda_1\rceil$, and thus

$$0\le\tau_c\le\left\lceil\frac{1}{\lambda_1}\right\rceil. \tag{22}$$

By this observation, (21) can therefore be solved very efficiently by the following procedure.

1) Initially, set $k_a=1$.
2) For the given k_a, the optimum τ_c is obtained by

$$\tau_c^*=\min_{0\le\tau_c\le\left\lceil\frac{1}{\lambda_1}\right\rceil}\tau_c \tag{23}$$

such that constraints (i) and (ii) in (21) are satisfied.

a) If τ_c^* can be obtained, the optimum k_a (denoted by k_a^*) is the present value. Therefore, optimization is reached by $\tau_c^*k_a^*$.
b) Otherwise, if $k_a<K$, set $k_a=k_a+1$ and repeat Step 2.

The complexity of our resource-optimal scheme is $\mathcal{O}(\lceil 1/\lambda_1\rceil)$, which is extremely applicable.

C. True End-to-End Packet-Forwarding Times

The expected values of the true end-to-end packet-forwarding time of a voice packet by leveraging k_a paths (i.e., $\bar{\Theta}_c$) and the true end-to-end packet-forwarding time of a bulk of video packets by leveraging k_a paths (i.e., $\bar{\Theta}_v$) suffering from different levels of link unavailability are still unclear. In the following, the detailed calculation for $\bar{\Theta}_c$ is provided. Denote $p_{k,f} \equiv \Pr\{\Theta_c^k = f\}$, $\bar{\Theta}_c$ is given by

$$\begin{aligned}\bar{\Theta}_c &= \mathbb{E}\left[\min\left\{\Theta_c^1, \ldots, \Theta_c^{k_a}\right\}\right] \\ &= \sum_{g=1}^{\infty} \Pr\left\{\min\left\{\Theta_c^1, \ldots, \Theta_c^{k_a}\right\} \geq g\right\} \\ &= \sum_{g=1}^{\infty} \Pr\left\{\Theta_c^1 \geq g, \Theta_c^2 \geq g, \ldots, \Theta_c^{k_a} \geq g\right\} \\ &= \sum_{g=1}^{\infty} \left[\sum_{f=g}^{\infty} p_{1,f} \times \cdots \times \sum_{f=g}^{\infty} p_{k_a,f}\right] \\ &= \sum_{g=1}^{\tau_c} \left[\sum_{f=g}^{\tau_c} p_{1,f} \times \cdots \times \sum_{f=g}^{\tau_c} p_{k_a,f}\right]. \end{aligned} \tag{24}$$

To derive $p_{k,f}$, two conditions shall be considered: 1) The true end-to-end packet-forwarding time of packet transmissions via the kth path exceeds τ_c with probability Φ_k; and 2) the true end-to-end packet-forwarding time of packet transmissions via the kth path does not exceed τ_c with probability $1 - \Phi_k$. Thus, $p_{k,f}$ can be expressed as

$$p_{k,f} = \Phi_k \Upsilon(f) + (1 - \Phi_k)\Gamma(k, f). \tag{25}$$

For 1), since a voice packet is only allocated by τ_c subframes, if the packet transmission violates the jitter constraint, then $f = \tau_c$, and $p_{k,f} = 1$. We therefore have

$$\Upsilon(f) = \begin{cases} 1, & \text{if } f = \tau_c \\ 0, & \text{otherwise.} \end{cases} \tag{26}$$

For 2), denote $S_{l,k}$ as the number of subframes to deliver a voice packet through the lth link of the kth path (the number of subframes that transmissions shall be suspended is not counted). It at least requires $\sum_{l=1}^{L_k} S_{l,k}$ subframes to deliver the packet via the kth path with L_k links, and therefore, $p_{k,f} = 0$ if $f < \sum_{l=1}^{L_k} S_{l,k}$ and $f > \tau_c$. If $\sum_{l=1}^{L_k} S_{l,k} \leq f \leq \tau_c$, $\Pr\{\Theta_k = f | \sum_{l=1}^{L_k} S_{l,k} \leq f \leq \tau_c\}$ is given by

$$\Omega_k = \prod_{l=1}^{L_k} \left\{ \sum_{r_{l,k}=0}^{f-\sum_{l=1}^{L_k} S_{l,k} - r_{l-1,k}} \binom{S_{l,k} - 1 - r_{l,k}}{r_{l,k}} \cdot (1 - \varphi_{l,k})^{r_{l,k}} (\varphi_{l,k})^{S_{l,k}} \right\}. \tag{27}$$

We consequently obtain

$$\Gamma(k, f) = \begin{cases} \Omega_k, & \text{if } \sum_{l=1}^{L_k} S_{l,k} \leq f \leq \tau_c \\ 0, & \text{otherwise.} \end{cases} \tag{28}$$

Finally, Φ_k is given by

$$\Phi_k = \Phi_{1,k} + \sum_{f=1}^{L_k - 1} \left(\prod_{g=1}^{f} (1 - \Phi_{g,k}) \right) \Phi_{f+1,k} \tag{29}$$

where $\Phi_{l,k}$ is the probability that the packet transmission via the lth link of the kth path violates the maximum tolerable jitter constraint, i.e.,

$$\Phi_{l,k} = \sum_{r=\tau'_{l,k} - S_{l,k} + 1}^{\tau'_{l,k}} \binom{\tau'_{l,k}}{r} (1 - \varphi_{l,k})^r (\varphi_{l,k})^{\tau'_{l,k} - r} \tag{30}$$

and $\tau'_{l,k}$ is the number of residue subframes before τ_c is expired. Thus, by (24)–(30), $\bar{\Theta}_c$ can be obtained.

By a similar method, $\bar{\Theta}_v$ can be obtained as well.

IV. Resource-Optimal Scheme for General Non-Realtime File Sources

As mentioned in Section III, there are two optimizations. First, each GNF source optimizes its path selection decision based on $0 \preceq \mathbf{Q} = [q_1, \ldots, q_{k_a}] \preceq 1$ broadcasted by the heterogeneous CRANs. Second, the heterogeneous CRANs then optimize $\mathbf{Q} = [q_1, \ldots, q_{k_a}]$ to control the path selection of all GNF sources. We should achieve these two optimizations in this section.

A. Problem Formulation and Analysis

The major goal of the resource-optimal framework is to fully utilize the minimum number of communication paths (i.e., k_a^*) to provide latency guarantees of n_c voice and n_v video sources. For this purpose, at each packet-forwarding period for GNF sources (as specified in Step 4 of the operation principle), if k_a^* communication paths are utilized, a maximum of k_a^* packets can be delivery via each of k_a^* communication paths. If the number of active GNF sources (the GNF sources with packets needing to be forwarded) is larger than k_a^*, then some GNF sources need to wait for the next packet-forwarding opportunity. Based on Step 4 of the operation principle of our scheme, the resource-optimal design for GNF sources can be mathematically formulated.

Optimization 3: Denote N_k as the set of GNF sources selecting the kth path and denote $\|N_k\|$ as the cardinality of N_k. The resource-optimal design for GNF sources, given that k_a^* communication paths are utilized, can be mathematically formulated by

$$\begin{aligned} \min_{q_1, \ldots, q_{k_a^*}} \quad & \max\left(\|N_1\|, \|N_2\|, \ldots, \|N_{k_a^*}\|\right) \\ \text{s.t.} \quad & \text{(i) } 0 \preceq \mathbf{Q} = [q_1, \ldots, q_{k_a^*}] \preceq 1 \\ & \text{(ii) } \|N_k\| q_k \leq 1, \text{ for } k = 1, \ldots, k_a^*. \end{aligned} \tag{31}$$

In (31), a set of *barrier parameter* $\mathbf{Q}$ is announced by the heterogeneous CRANs at the beginning of each packet-forwarding period for GNF sources. From Step 4 of the operation principle, if a GNF source selects the kth communication path for packet forwarding, this GNF source can transmit only a packet with probability $q' \leq q_k$ at the current packet-forwarding period. Therefore, the traffic load on the kth communication path can be optimally balanced over the time domain by designing q_k such that $\|N_k\|q_k \to 1$ while stability constraint (ii) is satisfied. Furthermore, $\mathbf{Q}$ not only controls the traffic load balance over the time domain but also reveals the congestion level on each of k_a^* communication paths. That is, a small value of q_k indicates that the kth communication path is overcongested, whereas a large value of q_k reveals that the kth communication path is underutilized. Given the announced $\mathbf{Q}$, each GNF source is able to autonomously select an appropriate communication path from k_a^* to forward its packet. Consequently, the traffic load can be optimally balanced over the spatial domain by the autonomous path selection in each GNF source. Under this framework, it is known that the number of GNF sources selecting the kth communication path (that is, $\|N_k\|$) for $k = 1, \ldots, k_a^*$ is also controlled by $\mathbf{Q}$. The time- and spatial-domain resource utilization can thus be jointly optimized via the optimum control of $\mathbf{Q}$.

From the perspective of engineering, constraint (ii) in (31) can be further relaxed to

$$\|N_k\|q_k \leq 1 + \varsigma, \text{ for } k = 1, \ldots, k_a^* \tag{32}$$

and the performance can still approach the optimum if ς is sufficiently small. If the path selection strategy in each GNF source is deterministic (that is, a GNF source only selects one specific communication path from k_a^* without influence by $\mathbf{Q}$), then $\|N_k\|$ for $k = 1, \ldots, k_a^*$ are deterministic. In this case, a feasible solution can be easily found as $q_k = 1/\|N_k\|$ such that the expected value of the number of GNF sources selecting the kth path is $\|N_k\|q_k = 1$. However, in the practical condition, the number of total active GNF sources n_a is too large to be known by each GNF sources, which makes $\|N_k\|$ unknown to each GNF source. In the following propositions, we show that the optimum communication path selection strategy for each GNF source is a statistical (mixed) strategy, instead of a deterministic (pure) strategy.

Proposition 1: Generally considering that $q_1 \geq q_2 \geq \cdots \geq q_{k_a^*} > 0$ are received by the nth GNF source (otherwise, these probabilities can be resorted), the nth GNF source shall adopt a mixed (statistical) strategy of

$$\beta_n(q_1, q_2, \ldots, q_{k_a^*}) = \left[q'_{n,1}, q'_{n,2}, \ldots, q'_{n,k_a^*}\right], \text{ where}$$
$$q'_{n,1} \geq q'_{n,2} \geq \cdots \geq q'_{n,k_a^*} > 0 \text{ and } \sum_{k=1}^{k_a^*} q'_{n,k} = 1 \tag{33}$$

where $q'_{n,k}$ is the probability that the nth GNF source selects the kth communication path.

Proof: We first consider the case of $\mathbf{Q}$ with only two elements, e.g., q_a and q_b, and $q_a > q_b$. Since the number of total active GNF sources n_a is unknown by each GNF source, all GNF sources receiving q_a and q_b adopt the same strategy, and it is common knowledge for all GNF sources. If the GNF source adopts the strategy with $q'_{n,a} < q'_{n,b}$, the bth path with severe congestion suffers from even more severe congestion since all GNF sources receiving q_a and q_b attempt to send packets via the bth path, whereas the slight congestion of the ath path is even eased. Therefore, adopting $q'_{n,a} < q'_{n,b}$ may not alleviate the congestion. Thus, the GNF source tends to change its strategy. On the other hand, if $q'_{n,a} > q'_{n,b}$ is adopted, the GNF source cannot choose a better strategy to further improve the performance (since the number of GNF sources adopting the same strategy is unknown to each GNF source). Thus, the GNF source may stay on this strategy. If $q_a = q_b$, the consequence of selecting the ath path and the bth path is equivalent. Thus, $q'_{n,a} = q'_{n,b}$ is adopted. Therefore, if $q_a \geq q_b$, the GNF source shall adopt $q'_{n,a} \geq q'_{n,b}$. This result can be extended to the case of $\mathbf{Q}$ with an arbitrary number of elements by pairwise arguments among all elements. Next, we prove the mixed strategy. If $q_a > q_b$ and $q'_{n,a} > q'_{n,b}$ while $q'_{n,b} = 0$, then $q_{n,a} = 1$ for the GNF source and other GNF sources receiving $q_a > q_b$ (that is, the pure strategy). If the pure strategy is adopted, although congestion in the bth path can be relaxed, congestion in the ath path may be extremely worse. Thus, the GNF source and other GNF sources tend to change their strategies. Thus, $q'_{n,a} > q'_{n,b} > 0$ shall be adopted, which suggests a mixed strategy. ■

Proposition 1 shows that the optimum path selection strategy is a mixed strategy. To satisfy (33), the general form of the strategy for each GNF source shall be as follows.

Proposition 2: Upon receiving $\mathbf{Q} = \{q_1, q_2, \ldots, q_{k_a^*}\}$, the nth GNF source adopts the strategy, i.e.,

$$\beta_n(q_1, q_2, \ldots, q_{k_a^*}) = \left[\frac{q_1}{\sum_{k=1}^{k_a^*} q_k} + \theta_1, \frac{q_2}{\sum_{k=1}^{k_a^*} q_k} + \theta_2, \ldots, \frac{q_{k_a^*}}{\sum_{k=1}^{k_a^*} q_k} + \theta_{k_a^*}\right], \sum_{k=1}^{k_a^*} \theta_k = 0. \tag{34}$$

Since strategies satisfying the general form of (34) are all equivalent, we can specify the strategy as

$$\beta_n(q_1, q_2, \ldots, q_{k_a^*}) = \left[q'_{n,1} = \frac{q_1}{\sum_{k=1}^{k_a^*} q_k}, q'_{n,2} = \frac{q_2}{\sum_{k=1}^{k_a^*} q_k}, \ldots, q'_{n,k_a^*} = \frac{q_{k_a^*}}{\sum_{k=1}^{k_a^*} q_k}\right]. \tag{35}$$

Given the optimum strategy on the selection of the communication path in (35), $\|N_k\|$ for all k can be obtained by

$$\|N_k\| = \sum_{n=1}^{n_a} q'_{n,k} \text{ for } k = 1, \ldots, k_a^*. \tag{36}$$

After providing the optimum strategy for each GNF source, the heterogeneous CRANs thus can control $\mathbf{Q}$ to jointly achieve the optimum traffic load balance over the time and spatial domains.

Optimization 4 (Reformulation of Optimization 3): The optimization of the resource-optimal scheme for GNF sources is mathematically formulated as

$$\min_{q_1,\ldots,q_{k_a^*}} \max\left(\sum_{n=1}^{n_a} q'_{n,1},\ldots,\sum_{n=1}^{n_a} q'_{n,k_a^*}\right)$$
$$\text{s.t.} \quad \text{(i)}\ 0 \preceq \mathbf{Q} = [q_1,\ldots,q_{k_a^*}] \preceq 1$$
$$\text{(ii)} \begin{cases} \sum_{n=1}^{n_a} q'_{n,1} q_1 \le 1+\varsigma \\ \quad\vdots \\ \sum_{n=1}^{n_a} q'_{n,k_a^*} q_{k_a^*} \le 1+\varsigma. \end{cases} \tag{37}$$

B. Optimization of the Design

To solve (37), we propose the following algorithm to optimize $\mathbf{Q}$.

In Algorithm 1, iterations stop when the difference of results between two successive iterations is not larger than ϵ. That is, $|\|\hat{N}_k\| - \|N_k^*\|| \le \epsilon$ for all k in "Row 4" of Algorithm 1. We will show, in the following section, that the number of iterations is extremely small, although ϵ is set to a small value of 10^{-4}.

Algorithm 1. Optimum Control of $\mathbf{Q}$

1: Initially, set $\|N_1^*\| = \|N_2^*\| = \cdots = \|N_{k_a^*}^*\| = n_a/k_a^*$
2: Set $q_k = 1/\|N_k^*\|$ for all k.
3: Set $\|\hat{N}_k\| = \sum_{n=1}^{n_a} q'_{n,k}$ for all k.
4: **while** $|\|\hat{N}_1\| - \|N_1^*\|| > \epsilon$ or $|\|\hat{N}_2\| - \|N_2^*\|| > \epsilon$ or $\ldots$ or $|\|\hat{N}_{k_a^*}\| - \|N_{k_a^*}^*\|| > \epsilon$ **do**
5: Set $\|N_k^*\| = \|\hat{N}_k\|$ for all k.
6: Set $q_k = 1/\|N_k^*\|$ for all k.
7: Set $\|\hat{N}_k\| = \sum_{n=1}^{n_a} q'_{n,k}$ for all k.
8: **end while**
9: **Output**: $\mathbf{Q} = [q_1, q_2, \ldots, q_{k_a^*}]$

Corollary 1: The output of Algorithm 1 converges to the optimum.

Proof: We consider the two-path case, and such a case can be easily extended to the k_a^*-path case. We first prove the convergence of Algorithm 1. Denote $\|N_1^{(t)}\|$ and $\|N_2^{(t)}\|$ as the numbers of GNF sources that may select the first path and the second path after the tth iteration. Therefore

$$\left\|N_1^{(t+1)}\right\| = \left\|N_1^{(t)}\right\| \cdot \frac{\frac{1}{\left(\left\|N_1^{(t-1)}\right\|+\nu\right)}}{\left(\frac{1}{\left(\left\|N_1^{(t-1)}\right\|+\varepsilon\right)}\right)+\left(\frac{1}{\left(\left\|N_2^{(t-1)}\right\|-\nu\right)}\right)} \tag{38}$$

or

$$\left\|N_1^{(t+1)}\right\| = \left\|N_1^{(t)}\right\| \cdot \frac{\frac{1}{\left(\left\|N_1^{(t-1)}\right\|-\nu\right)}}{\left(\frac{1}{\left(\left\|N_1^{(t-1)}\right\|-\varepsilon\right)}\right)+\left(\frac{1}{\left(\left\|N_2^{(t-1)}\right\|+\nu\right)}\right)} \tag{39}$$

where

$$\left\|N_1^{(t-1)}\right\| \pm \nu = \left\|N_1^{(t)}\right\| \tag{40}$$

$\nu \ge 0$ is the difference between $\|N_1^{(t-1)}\|$ and $\|N_1^{(t)}\|$. It is known that $\|N_1^{(t)}\|$ converges to a fixed value when $t \to \infty$ if

$$\frac{\frac{1}{\left(\left\|N_1^{(t-1)}\right\|+\nu\right)}}{\left(\frac{1}{\left(\left\|N_1^{(t-1)}\right\|+\nu\right)}\right)+\left(\frac{1}{\left(\left\|N_2^{(t-1)}\right\|-\nu\right)}\right)} = \frac{\left\|N_2^{(t-1)}\right\| - \nu}{\left\|N_1^{(t-1)}\right\| + \left\|N_2^{(t-1)}\right\|} < 1 \tag{41}$$

or

$$\frac{\frac{1}{\left(\left\|N_1^{(t-1)}\right\|-\nu\right)}}{\left(\frac{1}{\left(\left\|N_1^{(t-1)}\right\|-\nu\right)}\right)+\left(\frac{1}{\left(\left\|N_2^{(t-1)}\right\|+\nu\right)}\right)} = \frac{\left\|N_2^{(t-1)}\right\| + \nu}{\left\|N_1^{(t-1)}\right\| + \left\|N_2^{(t-1)}\right\|} < 1. \tag{42}$$

Therefore, ν is required to be less than $\|N_1^{(t-1)}\|$, and the following inequality:

$$\left|\left\|N_1^{(1)}\right\| - \left\|N_1^{(0)}\right\|\right| < \left\|N_1^{(0)}\right\| \tag{43}$$

shall be achieved. Similar to (38) and (39), it is known that

$$\left|\left\|N_2^{(1)}\right\| - \left\|N_2^{(0)}\right\|\right| < \left\|N_2^{(0)}\right\| \tag{44}$$

shall also be achieved. Since the initial setting of Algorithm 1 ($\|N_1^{(0)}\|$ and $\|N_2^{(0)}\|$) achieves the most balanced sharing of numbers of GNF sources among the first path and the second path, $|\|N_1^{(1)}\| - \|N_1^{(0)}\|| < \|N_1^{(0)}\|$ and $|\|N_2^{(1)}\| - \|N_2^{(0)}\|| < \|N_2^{(0)}\|$ can be achieved by adopting such an initiation. From (38) and (39), it is also known that $\varepsilon < \|N_1^{(t-1)}\|$ (and $\varepsilon < \|N_2^{(t-1)}\|$) is valid after the tth iteration for any t (ε decreases after each iteration). We thus complete the proof of convergence. Such convergence to a fixed value also suggests the optimality of $\mathbf{Q}$. Since iterations proceed until the stop rule is met, if there exists $\mathbf{Q}'$, $\mathbf{Q}' \neq \mathbf{Q}$, resulting in a better performance than that of $\mathbf{Q}$, then $\mathbf{Q}'$ cannot satisfy (ii), which suggests the optimality of $\mathbf{Q}$. ■

Please note that if a small ν can be achieved, it is known that

$$\frac{1}{\left\|N_k^{(t)}\right\|}\left(\left\|N_k^{(t)}\right\| + \nu\right) = 1 + \frac{\nu}{\left\|N_k^{(t)}\right\|} = 1 + \varsigma. \tag{45}$$

Therefore, $\varsigma < \nu$. Since $\nu \le \epsilon$, we have $\varsigma < \epsilon$. As a result, ς in (32) can be also acceptable.

TABLE II
SYSTEM PARAMETERS AND ASSUMPTIONS FOR SIMULATIONS

Parameters	Values/assumptions
Number of resource blocks in a subframe	100
Subframe length	1ms
Total number of communication paths	40
Number of links in each path	Uniformly distributed in [1,10]
Length of each link	Uniformly distributed in [50,500]m
Interference sources	WiFi stations
WiFi station deployment	Poisson Point Process with density $0.001/m^2$ [40]
TX power of WiFi station	20 dBm
TX power of RN	46 dBm
Packet size in LTE-A	142 Bytes
Unavailability on each link	$(1-\varphi_{l,k})$
Channel model on each link	Non-light-of-sight model in [39]
Number of active GNF sources	1000 to 5000
Arrival patent of voice source	voice model in [41]
Arrival patent of video source	MPEG4 model in [41]
Arrival patent of GNF source	FTP model in [39]

TABLE III
CHARACTERISTICS AND REQUIREMENTS OF VOICE AND VIDEO SOURCES (FROM [41])

	Voice1	Voice2	Voice3	Voice4	Voice5
λ[a]	0.05	0.04	0.03	0.03	0.03
δ	20ms	25ms	30ms	30ms	30ms
ε	0.02	0.02	0.02	0.02	0.02
	Video1	Video2	Video3	Video4	Video5
σ	69 pkt	38 pkt	48 pkt	66 pkt	61 pkt
ρ[b]	0.037	0.0012	0.091	0.037	0.0556
d	40ms	40ms	40ms	40ms	40ms
ξ	0.02	0.02	0.02	0.02	0.02

[a] λ is in the unit of subframes/packet arrival.
[b] ρ is in the unit of subframes/packet arrival.

V. PERFORMANCE EVALUATIONS

To evaluate the performance of our resource-optimal scheme for heterogeneous carrier communications in heterogeneous CRANs, we adopt LTE-A as a demonstration example. In LTE-A, a communication path from a data source to a destination station can be composed of multiple RNs to form a multihop connection (via multiple links). All these wireless links are vulnerable to link failures and intersystem interference. In this simulation, the number of total available communication paths K is 40, shared by all voice, video, and GNF sources. The number of links of the kth communication path, i.e., L_k, is uniformly distributed over [1, 10] to capture various cell deployments. For the performance evaluation, we adopt system parameters and assumptions of LTE-A in [39], as summarized in Table II.

A. Performance Evaluation for Voice/Video Transmissions

1) Latency Guarantee Provisioning: Before evaluating the performance of resource saving for our scheme, we shall first evaluate the essential requirement of latency guarantee provisioning. To generate voice and video packets, there are five classes of packet-generating rates and timing constraints for voice sources, which are listed in Table III. The specific data-generating rate of each voice source is randomly selected from these five classes [41]. For video sources, there are also five classes of packet-generating rates and timing constraints, according to the MPEG4 models defined in [41]. The jitter and delay violation probabilities for voice and video sources are 0.02, based on [39].

Table IV shows the simulation results of jitter and delay violation probabilities of five voice and five video sources by applying our resource-optimal scheme. In Table IV, results are shown in the form of (jitter or delay constraint violation probability, average availability of all links of the kth communication path, denoted by $\overline{\varphi}_k$). We can observe from Table IV that, by only utilizing one communication path $(k_a = 1)$, the timing constraints of five voice and five video sources can be satisfied when $\overline{\varphi}_k \geq 0.8$. Furthermore, if $k_a = 5$, five voice and five video sources can be supported, even under the severe interference condition of $\overline{\varphi}_k = 0.4$. These results show the effectiveness of our scheme to provide latency guarantees for voice and video transmissions using LAA in heterogeneous CRANs.

2) Resource Efficiency Evaluation: Table IV shows only the latency guarantee provisioning for real-time transmissions. To further show the optimality on resource efficiency, a comprehensive evaluation on the number of communication paths needed to satisfy the timing constraints of five voice and five video sources is demonstrated in Fig. 4. In this simulation, we particularly adopt the following two classes of remarkable state-of-the-art multipath transmission schemes as performance comparison benchmarks.

- **Conservative time-domain multipath transmission (CTMT)**: In the CTMT [35], replicates of each packet are forwarded through a given number of communication paths. For each packet, the packet-forwarding period is set to the maximum tolerable jitter/delay constraint. In other words, the packet is discarded if the timing constraint is expired while the packet is not delivered to the destination yet. Regardless of whether the packet is delivered to the destination on time, the subsequent packet transmission proceeds at the beginning of the next packet-forwarding period. The CTMT is a conservative scheme to maximize the capability of latency guarantee provisioning, which is particularly designed for the communication environment with severe interference.
- **Enhanced time-domain multipath transmission (ETMT)**: The ETMT is an improvement of the CTMT [42]. Similar to the CTMT, in the ETMT, replicates of each packet are forwarded through a given number of communication paths. However, in the ETMT, the packet-forwarding period is not fixed. Instead, if the timing constraint is expired while the packet is not delivered to the destination yet, the current packet-forwarding period ends, and the subsequent packet-forwarding period begins. If the packet is delivered to the destination before the timing constraint is expired, replicates of a feedback message are sent via the same multiple paths from the destination. If one of the replicated feedback messages is received by the source before the timing constraint is expired, the subsequent packet-forwarding period begins.

TABLE IV
SIMULATION RESULTS OF THE PROPOSED RESOURCE-OPTIMAL DESIGN ($\varepsilon = 0.02$ AND $\xi = 0.02$ FOR ALL VOICE AND VIDEO SOURCES)

Source	Voice1	Voice2	Voice3	Voice4	Voice5	Video1	Video2	Video3	Video4	Video5
k_a=1	(0.001,0.9), (0.008,0.8)	(0.001,0.9), (0.008,0.8)	(0.001,0.9), (0.008,0.8)	(0.002,0.9), (0.008,0.8)	(0.002,0.9), (0.008,0.8)	(0,0.9), (0,0.8)	(0,0.9), (0,0.8)	(0,0.9), (0,0.8)	(0,0.9), (0.006,0.8)	(0,0.9), (0.007,0.8)
k_a=3	(0,0.9), (0,0.8), (0,0.7), (0.002,0.6), (0.007,0.5)	(0,0.9), (0,0.8), (0,0.7), (0.002,0.6), (0.007,0.5)	(0,0.9), (0,0.8), (0,0.7), (0.002,0.6), (0.007,0.5),	(0,0.9), (0,0.8), (0,0.7), (0.002,0.6), (0.007,0.5)	(0,0.9), (0,0.8), (0,0.7), (0.002,0.6), (0.007,0.5)	(0,0.9), (0,0.8), (0,0.7), (0,0.6), (0.007,0.5)	(0,0.9), (0,0.8), (0,0.7), (0,0.6), (0.007,0.5),	(0,0.9), (0,0.8), (0,0.7), (0,0.6), (0.007,0.5)	(0,0.9), (0,0.8), (0,0.7), (0.001,0.6), (0.005,0.5)	(0,0.9), (0,0.8), (0,0.7), (0.002,0.6), (0.006,0.5)
k_a=5	(0,0.9), (0,0.8), (0,0.7), (0,0.6), (0.001,0.5), (0.003,0.4)	(0,0.9), (0,0.8), (0,0.7), (0,0.6), (0.001,0.5), (0.004,0.4)	(0,0.9), (0,0.8), (0,0.7), (0,0.6), (0.001,0.5), (0.004,0.4)	(0,0.9), (0,0.8), (0,0.7), (0,0.6), (0.001,0.5), (0.004,0.4)	(0,0.9), (0,0.8), (0,0.7), (0,0.6), (0.001,0.5), (0.004,0.4)	(0,0.9), (0,0.8), (0,0.7), (0,0.6), (0,0.5), (0,0.4)	(0,0.9), (0,0.8), (0,0.7), (0,0.6), (0,0.5), (0,0.4)	(0,0.9), (0,0.8), (0,0.7), (0,0.6), (0,0.5), (0,0.4)	(0,0.9), (0,0.8), (0,0.7), (0,0.6), (0.001,0.5), (0.003,0.4)	(0,0.9), (0,0.8), (0,0.7), (0,0.6), (0.001,0.5), (0.003,0.4)

For $k_a = 1$, timing constraints can be satisfied when $\overline{\varphi}_k \geq 0.8$. For $k_a = 3$ and $k_a = 5$, timing constraints can be satisfied when $\overline{\varphi}_k \geq 0.5$ and $\overline{\varphi}_k \geq 0.4$, respectively.

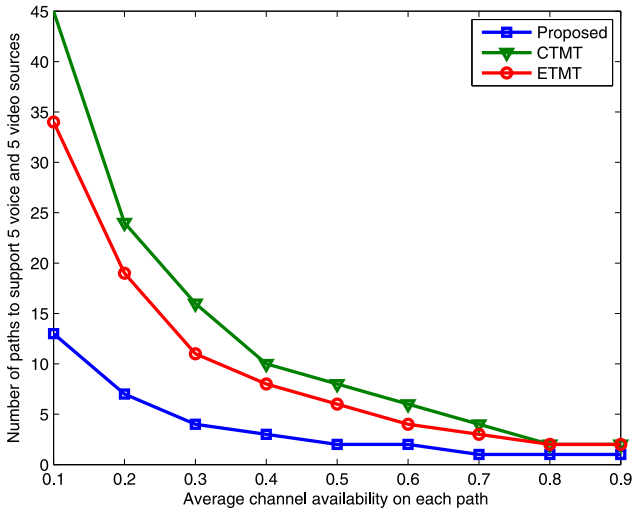

Fig. 4. Number of paths required to provide latency guarantees for five voice and five video sources under different channel availabilities.

Both the CTMT and ETMT are designed for wireless networks to combat intersystem interference. However, both of them do not optimize the radio resource utilization of multipath transmissions. Fig. 4 demonstrates that our resource-optimal scheme achieves outstanding resource efficiency as compared with those of the CTMT and ETMT. This result is not surprising. According to Table IV, if a large number of communication paths is utilized, the packet-forwarding time can be reduced. Consequently, we can set the packet delivery period to be stricter than the timing constraint to lead to a compact time-domain resource arrangement. However, since the time-domain resources for the CTMT may be wasted due to the conservative design, the CTMT needs more communication paths to support given numbers of voice and video sources. On the other hand, although the time-domain resource utilization is enhanced by ETMT, the ETMT still does not strike the optimal trade-off between latency guarantees and multipath utilization. The given results sufficiently demonstrate the outstanding resource utilization and the latency guarantee for our resource-optimal scheme.

B. Performance Evaluation for GNF Transmissions

1) Efficient Utilization of k_a^* Paths: To evaluate the efficiency on the utilization of k_a^* paths for GNF sources, we shall adopt the existing scheme in LTE-A [43] as a performance benchmark. In this legacy scheme, each GNF source selects one path from k_a^* communication paths in a deterministic fashion to forward its packets. Under this scheme, when GNF sources are typically with diverse packet arrival rates, the traffic load in a highly congested path cannot be shared by other paths with eased congestion. To capture such heterogeneous packet arrival rates in each GNF sources, it is considered that 60% among n_a GNF sources have a packet arrival rate that doubles the packet arrival rate of the remaining 40% among n_a GNF sources. In this simulation, we particularly focus on two classes of performance metrics: 1) the average time that a GNF packet waits for the selected communication path and the average successful probability that a GNF packet enjoys no wait; and 2) the worst case time that a GNF packet waits for the selected communication path and the worst case successful probability that a GNF packet enjoys no wait. Class 1) is of interest to the heterogeneous CRAN operators, whereas class 2) is of interest to the users of GNF sources.

In Fig. 5, we evaluate the performance of the average time and the worst case time that a GNF packet waits for the selected communication path. The average time is the performance averaged over all GNF sources, whereas the worst case performance is the largest waiting time among that of all GNF sources. We can observe the outstanding performance of our resource-optimal scheme, as compared with the existing LTE-A scheme. The reason for such performance is twofold. First, if multiple GNF sources select the same communication path simultaneously, only one packet can be served while other packets shall wait for the path returning to be idle. Under this case, our scheme facilitates GNF sources to select other communication paths with eased congestion. However, GNF sources adopting the existing scheme may suffer from continuous congestion. Second, the path selection of GNF sources is optimized to avoid subsequent congestion in other paths.

Next, the performance of the average successful probability and the worst case successful probability are shown in Fig. 6. Due to the optimization on the path selection, GNF sources

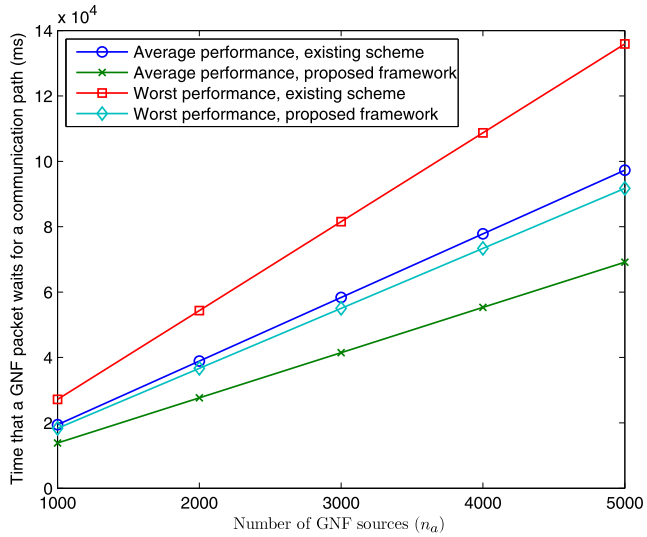

Fig. 5. Average and the worst case time that a GNF packet waits for a path. In this simulation, $\overline{\varphi}_k = 0.5$ and 10 paths are considered.

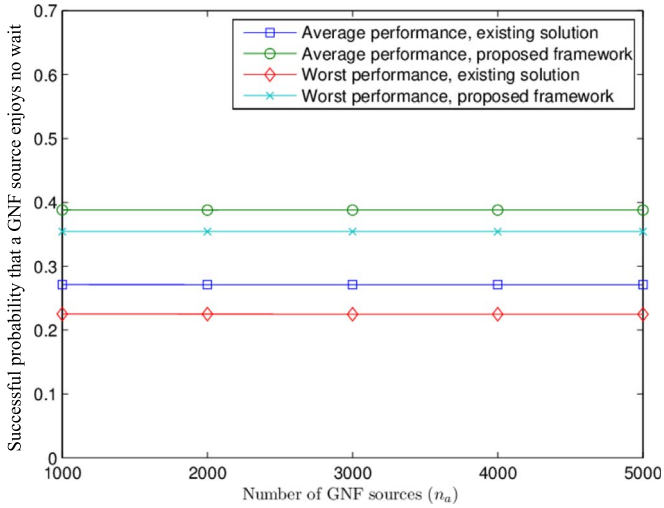

Fig. 6. Average and the worst case successful probabilities that a GNF source enjoys no wait. In this simulation, $\overline{\varphi}_k = 0.5$ and 10 communication paths are considered.

with our scheme enjoy a significant performance enhancement as compared with that of the existing scheme. It shows the efficiency on the utilization of k_a^* paths.

2) Complexity Evaluation: Finally, the major feature of the heterogeneous CRAN is the highly dynamic n_a. To support mobile users, the major concern lies in the complexity of Algorithm 1 to reach the optimum. For this concern, the number of iterations of Algorithm 1 to reach the optimum is evaluated in Fig. 7. We can observe that although ϵ (that is, the gap between the present result and the optimum result) is set to an extremely low value of 10^{-4}, there are only two to six iterations in Algorithm 1. This result demonstrates extreme complexity efficiency to quickly respond to the mobile environment.

VI. Conclusion

State-of-the-art cellular networks adopt radio resource scheduling and allocation to provide performance guarantees,

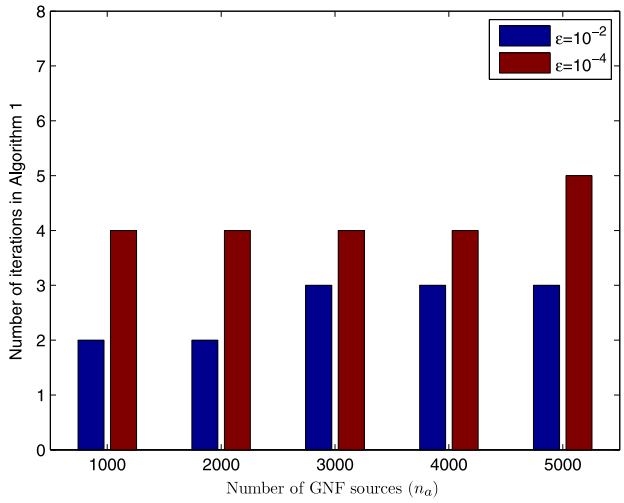

Fig. 7. Number of iterations in Algorithm 1 to reach the convergence.

which, however, leads to large uncertainty on the amount of available radio resources to adopt LAA. In this paper, our resource-optimal scheme creates a new design paradigm of using the minimum amount of replicated radio resources to optimally compensate unreliable communications. By successfully providing latency guarantees for real-time applications and maximized throughput for non-real-time applications, our scheme jointly achieves optimum resource efficiency and computation efficiency for LAA.

References

[1] S.-Y. Lien, K.-C. Chen, and Y. Lin, "Toward ubiquitous massive accesses in 3GPP machine-to-machine communications," *IEEE Commun. Mag.*, vol. 49, no. 4, pp. 66–74, Apr. 2011.

[2] K.-C. Chen and S.-Y. Lien, "Machine-to-machine communications: Technologies and challenges," *Ad Hoc Netw.*, vol. 18, pp. 3–23, Jul. 2014.

[3] A. Lo, L. Yee, and M. Jacobsson, "A cellular-centric service architecture for machine-to-machine (M2M) communications," *IEEE Commun. Mag.*, vol. 20, no. 5, pp. 143–151, Oct. 2013.

[4] J. G. Andrews *et al.*, "What will 5G be?" *IEEE J. Sel. Areas Commun.*, vol. 32, no. 6, pp. 1065–1082, Jul. 2014.

[5] S. Chen and J. Zhao, "The requirements, challenges, and technologies for 5G of terrestrial mobile telecommunication," *IEEE Commun. Mag.*, vol. 52, no. 5, pp. 36–43, May 2014.

[6] F. M. Abinader, E. P. L. de Almeida, F. S. Chaves, and A. M. Cavalcante, "Enabling the coexistence of LTE and Wi-Fi in unlicensed bands," *IEEE Commun. Mag.*, vol. 52, no. 11, pp. 54–61, Nov. 2014.

[7] "Feasibility study on licensed-assisted access to unlicensed spectrum," 3rd Generation Partnership Project, Sophia Antipolis Cedex, France, 3GPP TR 36.889 V13.0.0, Jun. 2015.

[8] N. Rupasinghe and I. Guvenc, "Licensed-assisted access for WiFi-LTE coexistence in the unlicensed spectrum," in *Proc. IEEE GLOBECOM Workshop*, 2014, pp. 894–899.

[9] C. Chen, R. Ratasuk, and A. Ghosh, "Downlink performance analysis of LTE and WiFi coexistence in unlicensed bands with a simple listen-before-talk scheme," in *Proc. IEEE VTC—Spring*, 2015, pp. 1–5.

[10] S.-Y. Lien and Y. J. Wang, "To random access or schedule? optimum 3GPP licensed-assisted access for machine-to-machine communications," in *Proc. IEEE QSHINE*, 2015, pp. 5–10.

[11] T.-E. Wu, K.-C. Chen, and D.-J. Deng, "Quality of experience in dense CSMA networks," in *Proc. IEEE ICC Workshop*, 2015, pp. 1759–1764.

[12] M. Peng, S. Yan, and H. V. Poor, "Ergodic capacity analysis of remote radio head associations in cloud radio access networks," *IEEE Wireless Commun. Lett.*, vol. 3, no. 4, pp. 365–368, Aug. 2014.

[13] C. Liu *et al.*, "A novel multi-service small-cell cloud radio access network for mobile backhaul and computing based on radio-over-fiber technologies," *J. Lightw. Technol.*, vol. 31, no. 17, pp. 2869–2875, Sep. 2013.

[14] Y. Shi, J. Zhang, and K. B. Letaief, "Group sparse beamforming for green Cloud-RAN," *IEEE Trans. Wireless Commun.*, vol. 13, no. 5, pp. 2809–2823, May 2014.
[15] J. You, Z. Zhong, G. Wang, and B. Ai, "Security and reliability performance analysis for cloud radio access networks with channel estimation errors," *IEEE Access*, vol. 2, pp. 1348–1358, Nov. 2014.
[16] C.-L. I *et al.*, "Recent progress on C-RAN centralization and cloudification," *IEEE Access*, vol. 2, pp. 1030–1039, Oct. 2013.
[17] P. Rost *et al.*, "Cloud technologies for flexible 5G radio access networks," *IEEE Commun. Mag.*, vol. 52, no. 5, pp. 68–76, May 2014.
[18] M. Peng, Y. Liu, D. Wei, W. Wang, and H.-H. Chen, "Hierarchical cooperative relay based heterogeneous networks," *J. Lightw. Technol.*, vol. 18, no. 3, pp. 48–56, Jun. 2011.
[19] S.-Y. Lien, Y.-Y. Lin, and K.-C. Chen, "Cognitive and game-theoretical radio resource management for autonomous femtocells with QoS guarantees," *IEEE Trans. Wireless Commun.*, vol. 10, no. 7, pp. 2196–2206, Jul. 2011.
[20] S.-M. Cheng, S.-Y. Lien, F.-S. Chu, and K.-C. Chen, "On exploiting cognitive radio to mitigate interference in macro/femto heterogeneous networks," *IEEE Wireless Commun. Mag.*, vol. 18, no. 3, pp. 40–47, Jun. 2011.
[21] Y. L. Lee, T. C. Chuah, J. Loo, and A. Vinel, "Recent advances in radio resource management for heterogeneous LTE/LTE-A networks," *IEEE Commun. Surveys Tuts.*, vol. 16, no. 4, pp. 2142–2180, Nov. 2014.
[22] A. Damnjanovic *et al.*, "A survey on 3GPP heterogeneous networks," *IEEE Wireless Commun. Mag.*, vol. 18, no. 3, pp. 10–21, Jun. 2011.
[23] J.-S. Lin and K.-T. Feng, "Femtocell access strategies in heterogeneous networks using a game theoretical framework," *IEEE Trans. Wireless Commun.*, vol. 13, no. 3, pp. 1208–1221, Mar. 2014.
[24] S. Barbarossa, S. Sardellitti, and P. D. Lorenzo, "Communicating while computing: Distributed mobile cloud computing over 5G heterogeneous networks," *IEEE Signal Process. Mag.*, vol. 31, no. 6, pp. 45–55, Nov. 2014.
[25] "Design of LAA-WiFi hidden terminal mitigation," Inst. Inf. Ind., 3rd Generation Partnership Project, Sophia Antipolis Cedex, France, 3GPP RAN1 80bis R1-151972, Apr. 2015.
[26] "Impact of LAA transmission gap on the hidden node problem for downlink LAA-WiFi coexistence," Cisco Systems, Diegem, Belgium, 3GPP RAN1 81 R1-153337, May 2015.
[27] "On hidden node aspects of LAA," Sony Corp., Tokyo, Japan, 3GPP RAN1 81 R1-153088, May 2015.
[28] "Broadband Radio Access Networks (BRAN): 5 GHz high performance RLAN," Eur. Telecommun. Std. Inst. (ETSI), Sophia Antipolis Cedex, France, ETSI EN 301 893 V1.8.0, Jan. 2015.
[29] L. Stanyer, *Modern Political Communications: Mediated Politics In Uncertain Terms*. Cambridge, U.K.: Polity Press, 2007.
[30] V. Kone, L. Yang, X. Yang, B. Y. Zhao, and H. Zheng, "The effectiveness of oppoutunistic spectrum access: A measurement study," *IEEE/ACM Trans. Netw.*, vol. 20, no. 6, pp. 2005–2016, Dec. 2012.
[31] Q. Liu, S. Zhou, and G. B. Giannakis, "Queueing with adaptive modulation and coding over wireless links: Cross-layer analysis and design," *IEEE Trans. Wireless Commun.*, vol. 4, no. 3, pp. 1142–1153, May 2005.
[32] Q. Liu, S. Zhou, and G. B. Giannakis, "Cross-layer combining of adaptive modulation and coding with truncated ARQ over wireless links," *IEEE Trans. Wireless Commun.*, vol. 3, no. 5, pp. 1746–1755, Sep. 2004.
[33] D. Wu and R. Negi, "Effective capacity: A wireless link model for support of quality of service," *IEEE Trans. Wireless Commun.*, vol. 12, no. 4, pp. 630–643, Jul. 2003.
[34] C.-S. Chang, K.-C. Chen, M.-Y. You, and J.-F. Chang, "Guaranteed quality-of-service wireless access to ATM networks," *IEEE J. Sel. Areas Commun.*, vol. 15, no. 1, pp. 106–118, Jan. 1997.
[35] P.-Y. Chen, W. C. Ao, and K.-C. Chen, "Rate–delay enhanced multipath transmission sheme via network coding in multihop networks," *IEEE Commun. Lett.*, vol. 16, no. 3, pp. 281–283, Mar. 2012.
[36] W. C. Ao, P.-Y. Chen, and K.-C. Chen, "Rate–reliability–delay tradeoff of multipath transmission using network coding," *IEEE Trans. Veh. Technol.*, vol. 61, no. 5, pp. 2336–2342, Jun. 2012.
[37] I.-W. Lai, C.-H. Lee, and K.-C. Chen, "A virtual MIMO path-time code for cognitive ad hoc networks," *IEEE Commun. Lett.*, vol. 17, no. 1, pp. 4–7, Jan. 2013.
[38] I.-W. Lai, C.-L. Chen, C.-H. Lee, and K.-C. Chen, "End-to-end virtual MIMO transmission in ad hoc cognitive radio networks," *IEEE Trans. Wireless Commun.*, vol. 13, no. 1, pp. 330–341, Jan. 2014.
[39] "Further advancements for E-UTRA physical layer aspects," 3rd Generation Partnership Project, Sophia Antipolis Cedex, France, 3GPP TR 36.814 V9.0.0, Mar. 2010.
[40] J. G. Andrew, F. Baccelli, and R. K. Ganti, "A tractable approach to coverage and rate in cellular networks," *IEEE Trans. Wireless Commun.*, vol. 59, no. 11, pp. 3122–3134, Nov. 2011.
[41] R. Srinivasan, J. Zhuang, L. Jalloul, R. Novak, and J. Park, "IEEE 802.16m Evaluation Methodology Document (EMD)," 2008. [Online]. Available: http://www.ieee802.org/16/tgm/docs/80216m-08_004r2.pdf.
[42] W.-C. Ao and K.-C. Chen, "End-to-end HARQ in cognitive radio networks," in *Proc. IEEE WCNC*, 2010, pp. 1–6.
[43] "Access Class Barring and Overload Protection (ACBOP)," 3rd Generation Partnership Project, Sophia Antipolis Cedex, France, 3GPP TR 23.898 V7.0.0, Apr. 2005.

Shao-Yu Lien received the B.S. degree from the Department of Electrical Engineering, National Taiwan Ocean University, Keelung, Taiwan, in 2004; the M.S. degree from the Institute of Computer and Communication Engineering, National Cheng Kung University, Tainan, Taiwan, in 2006; and the Ph.D. degree from the Graduate Institute of Communication Engineering, National Taiwan University, Taipei, Taiwan, in 2011.

Since February 2013, he has been an Assistant Professor with the Department of Electronic Engineering, National Formosa University, Yunlin, Taiwan. His research interests include optimization techniques for communication networks.

Mr. Lien received the IEEE Communications Society Asia-Pacific Outstanding Paper Award 2014, the Scopus Young Researcher Award (issued by Elsevier) 2014, the URSI AP-RASC 2013 Young Scientist Award, and the IEEE International Conference on Communications 2010 Best Paper Award.

Shin-Ming Cheng (S'05–M'07) received the B.S. and Ph.D. degrees in computer science and information engineering from National Taiwan University, Taipei, Taiwan, in 2000 and 2007, respectively.

From 2007 to 2012, he was a Postdoctoral Research Fellow with the Graduate Institute of Communication Engineering, National Taiwan University. Since 2012, he has been an Assistant Professor with the Department of Computer Science and Information Engineering, National Taiwan University of Science and Technology, Taipei. His current research interests include mobile networks, wireless communication, information security, and complex networks.

Dr. Cheng received the IEEE Personal, Indoor and Mobile Radio Communications 2013 Best Paper Award and the 2014 Association for Computing Machinery Taipei/Taiwan Chapter K. T. Li Young Researcher Award.

Kwang-Cheng Chen (M'89–SM'94–F'07) received the B.S. degree from National Taiwan University, Taipei, Taiwan, in 1983 and the M.S. and Ph.D. degrees from the University of Maryland, College Park, MD, USA, in 1987 and 1989, respectively, all in electrical engineering.

From 1987 to 1998, he was with SSE, COMSAT, the IBM Thomas J. Watson Research Center, and National Tsing Hua University, working on mobile communications and networks. Since 1998, he has been with National Taiwan University. After serving as the Director of the Graduate Institute of Communication Engineering, Communication Research Center and the Associate Dean for Academic Affairs, he later became a Distinguished Professor with National Taiwan University and is currently visiting the Massachusetts Institute of Technology, Cambridge, MA, USA, for the period 2015–2016. His recent research interests include wireless communications, network science, and data science.

Dr. Chen is actively involved in the organization of various IEEE conferences as the General/Technical Program Committee Chair/Cochair, and has served on the Editorial Board of several IEEE journals. He also actively participates in and has contributed essential technology to various IEEE 802, Bluetooth, and Long-Term Evolution (LTE) and LTE-Advanced wireless standards. He has received a number of awards, such as the 2011 IEEE Communications Society (COMSOC) Wireless Communications Technical Committee Recognition Award, the 2014 IEEE Jack Neubauer Memorial Award, and the 2014 IEEE COMSOC Asia-Pacific Outstanding Paper Award.

Dong In Kim (S'89–M'91–SM'02) received the Ph.D. degree in electrical engineering from the University of Southern California, Los Angeles, CA, USA, in 1990.

He was a tenured Professor with the School of Engineering Science, Simon Fraser University, Burnaby, BC, Canada. Since 2007, he has been with Sungkyunkwan University, Suwon, Korea, where he is currently a Professor with the College of Information and Communication Engineering.

Dr. Kim has served as an Editor and a Founding Area Editor of Cross-Layer Design and Optimization for the IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS from 2002 to 2011. From 2008 to 2011, he served as the Co-Editor-in-Chief for the IEEE JOURNAL OF COMMUNICATIONS AND NETWORKS. From 2012 to 2015, he served as the Founding Editor-in-Chief for the IEEE WIRELESS COMMUNICATIONS LETTERS. From 2001 to 2014, he served as an Editor of "Spread Spectrum Transmission and Access" for the IEEE TRANSACTIONS ON COMMUNICATIONS and then as an Editor-at-Large for *Wireless Communication*. He was the first recipient of the National Research Foundation of Korea Engineering Research Center in Wireless Communications for Energy Harvesting Wireless Communications (2014–2021).