This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMC.2018.2815702, IEEE Transactions on Mobile Computing

Delay Guaranteed Network Association for Mobile Machines in Heterogeneous Cloud Radio Access Network

Shao-Chou Hung, Hsiang Hsu, Shin-Ming Cheng, Qimei Cui , Kwang-Cheng Chen

Abstract—In a heterogeneous cloud radio access network (H-CRAN), which consists of multiple access points (APs) providing smaller coverage and a high power node (HPN) providing ubiquitous coverage, mobile machines can connect to multiple APs and a HPN by coordinated multi-point transmission (CoMP) concurrently to achieve ultra-reliable and low-latency communication. However, the current network association (or priorly known as handovers), which only focuses on switching between two base stations, may not be an efficient scheme in the H-CRAN. In this paper, we innovate a proactive network association mechanism by taking CoMP into consideration under the H-CRAN architecture. We consider two scenarios under the H-CRAN architecture: with and without the assistance of the HPN in the network. By regarding APs/HPN in the H-CRAN as resources that allocated to mobile machines, a novel proactive network association concept is proposed, and then generalized from one-to-one to multiple-to-multiple case. With the assistance of *Lyapunov optimization theory*, *effective bandwidth and capacity theory*, we can prove that this proactive network association scheme can guarantee that the queueing delay performance and the delay violation probability can be both smaller than a corresponding upper bound. That is, both low-latency and ultra-reliable communication can be guaranteed. We also conduct experiments by using real trace from taxis movement data to verify the analytical results. Out results suggest the guidelines to design the proactive network association scheme in a H-CRAN.

Index Terms—Network association, Cloud radio access network, CoMP, Robotic communication, Autonomous vehicles, Machine-to-Machine communication, Vertical handover, Internet of Things, Ultra-reliable and low-latency communication

1 Introduction

The development of autonomous driving or robots has attracted interest due to its potential of improving traffic safety, efficiency, and information dissemination. Most autonomous vehicles (AVs), *e.g.*, Google Car [1], have been developed based on a perception system, including various on-board sensors and machine intelligence to maneuver along the streets with other vehicles. Nevertheless, the intelligence of individual AV can be further enhanced by the networking and computing infrastructures of the entire intelligent transportation systems (ITS). Due to the limitations of on-board perception sensors, driving safety and efficiency in holistic scope of ITS heavily rely on the reliable and low-latency wireless networking toward success control information exchanges [2].

Upcoming intelligent mobile machines (IMMs) including autonomous and smart vehicles, unmanned aerial vehicles, robots, *etc*, are expected to reach the amount similar to smart phones. With supporting sensor and information infrastructures, current wireless networking technology cannot support the traffic volume and corresponding performance requirements, particularly networking delay. Furthermore, the safety of ITS highly relies on ultra-reliable and low-latency communication among the vehicles and associated vehicular and mobile networks. The safety-related messages for reliable ITS demand strict networking requirements. According to [3], the delay performance of safety-related messages should be no more than $50 \sim 100ms$. For massive operation of autonomous vehicles, it is widely believed that a further stringent end-to-end latency in the order of $1ms$ is necessary [4]. To achieve ultra-reliable and low-latency communication in ITS, there are two major technology challenges to overcome: (1) Spectrum scarcity and (2) Network association of low delay guarantees.

- Spectrum scarcity: From GSM/GPSR, UMTS to LTE/LTE-A, data transmission rate has been enhanced to a million fold solely by connecting to a powerful widely base stations (BSs). With the development of physical layer technology such as MIMO, beamforming, it seems like that transmission rate has almost approached Shannon bound and cannot be improved largely [5]. To solve this challenge by small-cell ultra-dense networking, the architecture of heterogeneous cloud radio access network (H-CRAN) was proposed as shown in Fig. 1. In general, there are two major tiers of networks under H-CRAN archiecture. The first tier is composed of high power node (HPN), which traditionally can provide the ubiquitous services of the IMMs. The second tier is composed of a group of distributed low power APs in the service area of the HPN. By decreasing the distance between IMMs and APs, the spectrum efficiency and transmission rate can be successfully improved [6].
- Network association: The smaller transmission distance

- *Shao-Chou Hung and Hsiang Hsu are with the Graduate Institute of Communication Engineering, National Taiwan University, Taipei, Taiwan E-mail: d02942008@ntu.edu.tw and r03942046@ntu.edu.tw*
- *Shin-Ming Cheng is with National Taiwan University of Science and Technology, Taipei, Taiwan. E-mail: smcheng@mail.ntust.edu.tw*
- *Qimei Cui is with the National Engineering Laboratory for Mobile Network Technologies, Beijing University of Posts and Telecommunications., Beijing, China. E-mail: cuiqimei@bupt.edu.cn.*
- *Kwang-Cheng Chen is with the Department of Electrical Engineering, University of South Flordia, Tampa, Flordia. E-mail: kwangcheng@usf.edu.*

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMC.2018.2815702, IEEE Transactions on Mobile Computing

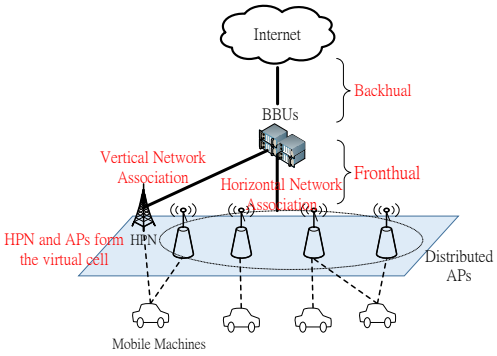

Fig. 1: H-CRAN based air interface architecture.

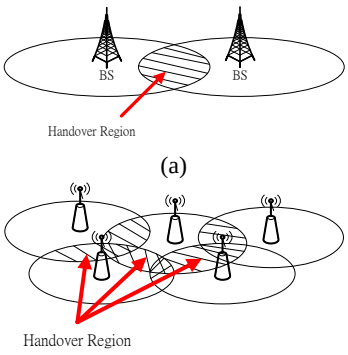

Fig. 2: a) The conventional handover in a homogeneous network. b) Frequent handover in the heterogeneous networks.

in H-CRAN architecture suffers from frequent network association (also known as user association or handover). As shown in Fig. 2, conventional network associations only happen at the edge of a BS coverage. However, under the H-CRAN architecture, there are many distributed APs and each of them has a smaller service region than the coverage of a conventional BS. The edge of the networks are any-where. Under this scenario, to prevent from the network being occupied by control signals, it is necessary to coordinate the small cell networks to execute a new handover scheme [7]. Therefore, the concept of *virtual cell* is proposed to solve this problem. It is achieved by connecting all distributed APs (or called remote radio head) and HPN with the Baseband Units (BBUs) to create a large cell virtually, and all the radio resources in this "large" cell is scheduled and allocated by utilizing the cloud computing technology [8]. In this case, the different APs are transparent to the IMMs. The number of network associations is successfully reduced.

In order to fully take advantage of the distributed APs, coordinated multi-point transmission (CoMP) has been considered as an important technique in H-CRAN [9] within virtual cell. In this case, all the distributed APs behave as the remote antennas, and are allowable to service one IMM simultaneously. The data collected from these remote antennas can be processed at BBU in a centralized way. It is intuitive for the IMMs to build more connections to the APs or the HPN to pursue much better transmission performance. However, the more connections with the APs, the larger number of signalling overheads might be incurred. For example, the signaling overhead due to synchronization among APs in the same virtual cell might become terrible burdens to meet the delay requirements [10]. On the other hand, connecting with all the available APs may incur unfairness issue. Consider that an IMM occupies all the available APs, a later coming IMM suffering from deep fading or severe interference may not obtain enough APs. The delay performance thus cannot be guaranteed. As a result, the number of utilized APs should also be optimized.

The goal of this paper is to design a delay-aware "vehicle"-centric approach to fully utilize the advantages of CoMP and simultaneously prevent the resulting control signal overheads to hurt the delay performance, such that ultra-low latency end-to-end wireless networking can be realized. A novel proactive network association mechanism is proposed to enable effective radio resource utilization by taking fairness and control signal overheads into consideration. We focus on guaranteeing the delay performance with only "enough" utilization number of APs and thus the control signal overheads can be minimized. In addition, comparing with the network association in a conventional one-tier network consisting of only BSs, network association in H-CRAN with two-tier architecture becomes more complicated. The IMMs might coordinate with the multiple APs (known as the horizontal network association) or connect to the HPN (known as vertical network association) and the APs concurrently. The transmission quality of the IMMs is highly relative to the number of APs being connected to and whether the IMMs have a connection to the HPN or not. Therefore, the network associations in H-CRAN become a process of allocation involved the distributed APs and HPN. In other words, we can regard these distributed APs and HPN as the limited cherish resource in H-CRAN and IMMs can proactively access such "resources" to improve their delay performance. Different from conventional approach, switching from one BS to another BS (one-to-one scenario), the proposed proactive network association thus becomes a multiple-to-multiple scenario, *i.e.*, IMMs switch from a set of APs to another set of APs.

We discuss two different scenarios in H-CRAN: with and without the assistance of the HPN. In the scenario without HPN, the IMMs cannot access to the HPN and only rely on the distributed APs. The goal is to design a proactive network association scheme that can minimize the number of the utilized APs by the IMMs but simultaneously guarantee the delay requirements. It is not only for decrease the complexity of decoding CoMP signal but also to decrease the burden on the backhaul network [11]. In the scenario with HPN, an IMM can make a decision whether to connect to HPN for better service or not. In this scenario, the utilization of HPN is minimized under the constraint of guaranteeing the delay performance. The first reason is to avoid additional information exchanges like authentication, which may further hurt the delay performance. Another reason is for keeping the infrastructures for emergency accident. Due to being able to provide ubiquitous service, HPNs should be reserved to those IMMs suffering from serious delay or emergencies like, car crash. Therefore, the design goal is to treat HPN as a supplementary infrastructure by minimizing utilization to achieve the delay guarantees.

With the high speed mobility, the number of available APs for an IMM changes faster than the conventional user equipments experiences. The conventional static optimization approach thus suffers from the out-of-date information and the performance cannot be optimized. Therefore, two time-dynamic optimization problems are formulated for two different scenarios to solve the

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMC.2018.2815702, IEEE Transactions on Mobile Computing

proactive network association in H-CRAN. The algorithms are also proposed to solve the corresponding dynamic optimization problems in the corresponding scenario. With the assistance of *Lyapunov optimization theory*, we can analyze the delay performance of the proposed algorithm. The proposed algorithm can also be proved to approach the best-tradeoff solution for the proposed dynamic optimization problem.

The rest of the paper is organized as follows: The related works are in Section 2. Section 3 describes the details of our model. In Section 4, we formulate the dynamic optimization for the proactive network association without the help of vertical association. The dynamic optimization for vertical association is described in Section 5. The delay violation probability is analyzed in Section 6. In Section 7, the design guideline for the horizontal and vertical network association are provided. The simulation results of the proposed algorithm are provided in Section 8.

2 Literature Reviews

In this section, we first revisit the literatures about data offloading in heterogeneous networks and CoMP clustering to reveal our unique contribution. Intuitively, the device makes a network association decision according to the radio link quality, *i.e.*, received signal strength (RSS) or signal-to-interference-plus-noise (SINR). Actually, this approach is also commonly used in the handover scheme of the traditional homogeneous networks [12]. However, with the amount of the IMMs increasing, solely selecting the network with the best link may not be a suitable strategy. For example, [13] points out that the users' quality of experiences (QoE) depends not solely on the SINR but also the other competitors and the corresponding allocated resource.

2.1 Data Offloading

To increase the performance of the whole systems, we can solve the network association from the viewpoints of loading balancing between different tiers of heterogeneous networks. That is, the users can be associated to multiple networks to decrease the loading of HPNs and fully utilize the unused resource in the small cells, and *vice versa*. Generally speaking, the current studies on data offloading focus on optimizing the system performance by a centralized [14]–[17] or a distributed [18]–[20] approach.

In [14], the load balancing scheme is designed based on the long-term throughput of users to find the best associations between HPNs and users. To decrease the network-wide average packet delay, the user association and corresponding resource allocation scheme are designed based on experiencing packet delay in [15]. In [16], the mobility model is taken into consideration. Based on the proposed mobility model, a Markovian-based approach is proposed to do data offloading. Energy consumption and network capacity are further improved. Data offloading gain in terms of delay performance is analyzed in [17]. In this works, the WiFi networks are regarded as the small cell networks utilizing the orthogonal radio resource to the HPN. Wifi networks are considered as small cell networks utilizing orthogonal radio resources to HPN. For distributed approaches, game theory is a popular tool to design a distributed strategy among different players (networks). In [18], a data offloading scheme based on the Stackelberg game and the corresponding efficient algorithm to find the best strategy are proposed. To measure the fairness between different network layers, a coalitional game is formed to encourage the cooperation between different layer networks in [19]. To fully utilize the unused capacity, an auction-based mechanism is proposed to encourage non-busy WiFis to release the resource and decrease the burden of HPN in [20].

2.2 CoMP Clustering

It is known of advantages to utilize CoMP in the heterogeneous networks. However, the success of CoMP relies on the coordination between all small APs including the synchronization issue [21], exchanges of channel state information [22], and additional signal processing for interference mitigation [23], *etc*. Consequently, in order to reduce these overhead, a question called CoMP clustering arise: how small a CoMP can be but still provide the major portion of the potential CoMP performance. User-centric approach is considered in [24], [25], where the clustering of the CoMP is dynamic according to the users. In [26], the user-centric approach is also proposed to maximize the average throughput of the network. Different from previous two, the limitation of the backhaul networks is considered. In [27], based on coalition game theory, a distributed clustering algorithm is proposed. The cluster size can automatically increase to the predefined cluster size and optimize the performance.

The previous works about data offloading assume that the IMMs can be served by only one HPN or AP. The IMMs are limited to switch from one HPN to another one. With the development of CoMP, the IMMs can connect to multiple APs simultaneously and a new network association should be developed. On the other hand, for the previous work about CoMP clustering, they focus on the optimize the performance with minimizing the size of CoMP network in the same tier. However, there still lacks method of CoMP clustering with assistance from HPN in a heterogeneous network. In this work, we tackle the network association from the viewpoint of minimizing the number of the utilized APs, *i.e.*, the size of CoMP. With vertical network association, a HPN can be regarded as an auxiliary tool. An IMM offloads data to a HPN when APs cannot satisfy service requirements. By this manner, we save more control signals to coordinate multiple APs to operate CoMP and the delay can be further improved.

We summarize our contribution as follows.

- A proactive horizontal network association is proposed to minimize the control signal cost of CoMP.
- A proactive vertical network association is proposed to fully integrate the APs and HPN networks.
- The proposed approach can only utilize "enough" APs or HPN to achieve the delay performance requirements.

3 System Model

3.1 The Cost of Network Connection

As mentioned before, the control signal overheads have a great impact on the delay performance. Therefore, it is necessary to know the time costs of horizontal and vertical associations carefully. We provide a brief discussion about what kind of time cost is needed in the following respectively. We refer to [28], [29] for the details numerical values of these time costs.

During the process of a horizontal network association, the following steps are required. First, it takes sensing delay (d_{sen}) to identify the available APs in its transmission region. Then, an IMM spends additional time to inform the core network about its connection requirement ($d_{\text{inform}}^{\text{IMM}}$). Subsequently, BBUs pool determines one or multiple appropriate APs to allocate a channel

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMC.2018.2815702, IEEE Transactions on Mobile Computing

TABLE 1: Glossary of Notations.

Notation	Description
J	Number of independent channels shared by APs.
λ_M	Distribution density of IMMs.
λ_{ap}	Distribution density of APs.
P_0	Transmission power of IMMs.
v	Velocity of IMMs.
θ	Non-outage $SINR$ threshold.
α	Path-loss exponent.
d	Distance between an IMM and an AP.
R	Radius of the transmission region of IMMs and APs.
G_j	Channel fading of an IMM in the jth channel.
G_{xj}	Channel fading of the xth IMMs as an interference source in jth channel.
p	Non-outage probability of each channel.
P_{av}	Probability of the AP being available for the arrival IMMs.
P_{vi}	Delay violation probability.
$a_A(t)$	In the scenario with vertical handover, $a_A(t)$ refers to the data go through APs.
P_{choose}	Probability that an AP is selected by an IMM.
P_{av}	Probability that not all J channels of an AP are occupied.
$N(t)$	Number of available APs at time slot t

to the IMM (d_{BBU}). If this allocated channel is the same with current one used by the IMM, the IMM just transmits data as previously. There is no further control information exchange needed between the IMM and the core network. The resulting total delay is

$$D_{\text{same}}^{\text{chan}} = d_{\text{sen}} + d_{\text{inform}}^{\text{IMM}} + d_{\text{BBU}}.$$

However, if the same channel in the newly designated AP has been allocated, this AP can immediately allocate the other channel to the IMM. In such case, the network needs to spend additional time to inform the IMM which channel is allocated to ($d_{\text{inform}}^{\text{Net}}$). Then, the IMM switches to the newly designated channel and resume the transmission after synchronization with multiple APs (d_{syn}) for CoMP. Then the resulting total delay is

$$D_{\text{diff}}^{\text{chan}} = d_{\text{sen}} + d_{\text{inform}}^{\text{IMM}} + d_{\text{BBU}} + d_{\text{inform}}^{\text{Net}} + d_{\text{syn}}.$$

Because $d_{\text{inform}}^{\text{Net}}$ and d_{syn} are two additional delays, to speed up association for low delay, it is better for the whole network to operate horizontal associations in the same channel as long as possible. Therefore, it is reasonable to assume that the networks always manage the horizontal association in the same channel.

An IMM can proceed vertical network association with a HPN. The vertical association involves further more complicated procedures between the IMM and the network. These additional procedures includes: (1) the sensing delay to find the existing HPN ($d_{\text{sen}}^{\text{ver}}$) (2) the IMM informs the core network the requirements of building a vertical connection ($d_{\text{inform}}^{\text{ver}}$) (3) the processing time in the BBU to coordinate the unoccupied channels (d_{HPN}) of the HPN (4) the time to inform the IMM about the allocated channel ($d_{\text{inform}}^{\text{Net}}$) (5) the synchronization time between the IMM and the HPN and APs ($d_{\text{syn}}^{\text{ver}}$). Therefore, the total delay D_{ver} is

$$D_{\text{ver}} = d_{\text{sen}}^{\text{ver}} + d_{\text{inform}}^{\text{ver}} + d_{\text{HPN}} + d_{\text{inform}}^{\text{Net}} + d_{\text{syn}}^{\text{ver}}.$$

On the other hand, due to the limited number of channels in a HPN, the BBUs may terminate the vertical connection with an IMM. This termination also introduces additional control signals to the network. To minimize the utilization rate of a HPN, it is not only the reason that these additional delay caused by vertical association but also these additional exchanges of control signals.

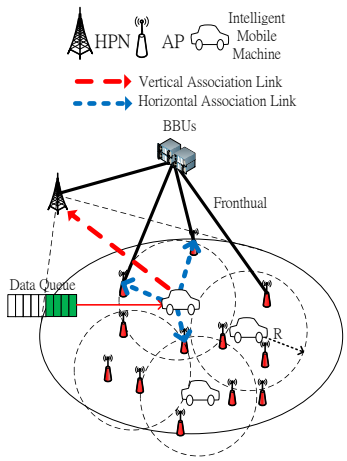

Fig. 3: Illustration of the network system architecture.

Therefore, it is necessary to design a new association scheme to decrease the utilization rate of *vertical association*.

3.2 Network Model

As shown in Fig. 3, the H-CRAN is composed of IMMs, a HPN, small APs and BBUs. The BBUs pool in the H-CRAN is in charge of collecting information and do the signal processing [9]. There are two possible paths in the air for the IMMs connecting to the network infrastructure. The first one is to build horizontal connections with the AP networks. We assume that there are J different channels shared by the AP networks. Each AP (or multiple APs with CoMP technique) can allocate one of the J channels to the IMMs in its service region. The IMMs can connect to the network with one or multiple APs through one of these J channels. The second one is a vertical connection with the HPN network. Here, we assume that the channel of horizontal and vertical air interface are orthogonal thus no cross-tier interference between the HPN and the APs networks.

To consider a general case, we take advantage of random modeling for the IMMs and the APs via the Poisson point process (PPP) model. The spatial distribution of the IMMs follows a PPP with density λ_M and its transmission power is denoted as P_0. Without loss of generality, we assume that each IMM is moving along the straight line with a velocity $v(meters/s)$ and different direction randomly during the need of network association or handover. The spatial distribution of the APs also follows PPP with density λ_{ap}. The IMMs can build the connections with the APs within a radius R through one or multiple of J channels. We assume that the packet can be transmitted successfully if the SIR_j value, the signal-to-interferenceratio in the jth channel, is larger than θ. Then the non-outage probability p_j in the jth channel is defined as

$$p_j \triangleq \mathbb{P}(SIR_j \geq \theta). \tag{1}$$

The signal-to-interference ratio SIR_j in the jth channel is

$$SIR_j \triangleq \frac{P_0 G_j R^{-\alpha}}{I_j} \tag{2}$$

where I_j is denoted as the total interference in the jth channel, α is the coefficient of path loss subject to environment, and G_j is the channel fading following exponential distribution with unit mean. Here, we ignore the effect of white noise due to strong interference from other IMMs.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMC.2018.2815702, IEEE Transactions on Mobile Computing

The interference I_j comes from the other IMMs also utilizing the jth channel. Here, we assume that all the APs randomly allocate a channel to each IMM. Therefore, the point process of the IMMs in the jth channels, denoted as Φ_j, is another PPP. Then, the interference in the jth channel can be expressed as

$$I_j = \sum_{x \in \Phi_j} P_0 G_{xj} R_x^{-\alpha}, \tag{3}$$

where G_{xj} is the channel fading from the xth IMM in the jth channel.

Here, we assume that all the IMMs choose the APs randomly. We also assume that, after an IMM connects to a set of APs, the BBUs pool or AP itself will allocates one channel to the IMM randomly. Therefore, the IMM's distribution density in the jth channel is $\lambda_j = \lambda_M / J$. Then the non-outage probability p_j of the transmission link in jth channel can be further expressed as [30]

$$\begin{aligned} p_j(d) &= \mathbb{P}\left(SIR_j \geq \theta\right) \\ &= \mathbb{E}_{I_j}\left[\mathbb{P}\left(G_j \geq \frac{\theta I_j}{P_0 d^{-\alpha}}\right)\right] \\ &= \exp\left[-\frac{\lambda_M}{J}\theta^{2/\alpha} d^2 \frac{2\pi^2}{\alpha \sin(2\pi/\alpha)}\right], \end{aligned} \tag{4}$$

where d is the distance between the IMM and AP. Due to the assumption that the IMMs choose the APs within radius R randomly, the distance between an IMM and an AP r is also a random variable with the distribution $r \sim 2r/R^2$. The non-outage probability p_j can be further expressed as

$$\begin{aligned} p &= \mathbb{E}_d(p_j(d)) \\ &= \int_{r=0}^{R} p_j(r) \frac{2r}{R^2} dr \\ &= \frac{1 - \exp\left(-\xi R^2\right)}{\xi R^2}, \end{aligned} \tag{5}$$

where $\xi = \frac{\lambda_M}{J}\theta^{2/\alpha}\frac{2\pi^2}{\alpha \sin(2\pi/\alpha)}$, which is actually the average number of available APs. We can find that the non-outage probability is the same in each channel. Therefore, non-outage probability in every channel is denoted as p, *i.e.*, $p_j = p, \forall j \in J$.

3.3 Queue Model for Device Mobility

Under the H-CRAN architecture with CoMP technique, data service rate of the IMMs depends on the available APs in the transmission region. The BBUs can allocate the available APs to the IMMs and thus control the data service rate. To describe the solution space of the proposed dynamic optimization and analyze the performance of the proposed scheme in the later section, it is necessary to find the probability distribution of the available APs around the IMMs. By the stationary characteristics of a homogeneous PPP, the statistics measured by a typical IMM at the origin is representative of all the others [31]. In the following, we consider a typical AP and IMM to represent all the others in the network.

3.3.1 Probability of APs Being Available

From the viewpoint of the AP, the IMMs enter into a circular transmission region centering at the AP with a radius R and are serviced by the AP until they leave the transmission region. Because the AP has only J channels, therefore, the number of the IMMs in the AP at each time slot can be modeled as a queue with J servers. Due to Poisson distribution of the IMMs, this queue can be regarded as a $M/G/J/J$ queue. J/J comes from the fact that each of J channels can serve only one IMM. The service time of an IMM is a duration of staying in the transmission region of the AP. The expected service time is $\frac{\pi R}{2v}$ as shown in Appendix B. The arrival rate of this queue is a rate that the IMM enters into the transmission region of the AP. It can be expressed as $2Rv\lambda_M P_{choose}$ as shown in Appendix B. The reason of P_{choose} is that the IMM may have multiple connectible APs simultaneously. We assume that the IMM randomly selects one of the connectible APs to establish a connection link. P_{choose} can be further expressed as

$$P_{choose} = \sum_{k=1}^{\infty} p_{n=k} \frac{1}{k}, \tag{6}$$

where $p_{n=k}$ is the probability that there are k APs in its transmission region of the IMM. It can be expressed as

$$\begin{aligned} p_{n=k} &= \mathbb{P}(\text{k possible APs} | k \geq 1) \\ &= \frac{e^{-\lambda_{ap}\pi R^2}(\lambda_{ap}\pi R^2)^k}{k!} \Bigg/ \sum_{i=1}^{\infty} \frac{e^{-\lambda_{ap}\pi R^2}(\lambda_{ap}\pi R^2)^n}{n!}, \end{aligned} \tag{7}$$

and $\frac{1}{k}$ in (6) comes from that the IMM randomly selects one AP among k connectible APs to establish a connection link.

To derive the probability of an AP being available, *i.e.*, not all the J channels are occupied by the IMMs, we need to know the probability distribution of the number of occupied server in the $M/G/J/J$ queue. According to queueing theory, it can be described by *Erlang B formula* [32]. We denote the probability of the AP being available as P_{av}, then it can be expressed as

$$\begin{aligned} P_{av} &= P(\text{not all server being occupied}) \\ &= 1 - \frac{\rho^J / J!}{\sum_{j=0}^{J} \rho^j / j!}, \end{aligned} \tag{8}$$

where $\rho = \pi R^2 \lambda_M P_{choose}$ is the utilization factor of a queue. Here, we need to note that it is an upper bound of the real probability of the APs being available because the IMMs may access multiple APs at the same time, which results in larger P_{choose} and thus smaller P_{av}.

3.3.2 Probability Distribution of Available APs for IMMs

In this subsection, we will derive the probability distribution of the number of available APs from the viewpoint of an IMM. We denote the number of available APs within the radius R centered at the IMM as $N(t)$. Due to high mobility, $N(t)$ is a random variable at each time slot t. The update rule of $N(t)$ can be expressed as

$$N(t+1) = \max[N(t) - N_l(t), 0] + N_a(t), \tag{9}$$

where $N_l(t)$ and $N_a(t)$ is the number of APs leaving and arriving the transmission region of the IMM. We can find that this is a classical form of a dynamic queue. Because there is no restriction on $N(t)$, the number of available APs in the transmission region of the IMM can be modeled as a $M/G/\infty$. With a velocity v, the expectation of $N_a(t)$ is $2Rv\lambda_{ap}P_{av}$ and the proof is similar to the one in Appendix B. The expectation service time of this queue, *i.e.*, the expected duration of an AP staying in the transmission region of the IMM is $\frac{\pi R}{2v}$, which can be derived by the similar steps in Appendix B. According to the transition behavior of a $M/G/\infty$ queue, the probability distribution of $N(t)$ is

$$\mathbb{P}(N(t) = n) = \frac{e^{-N_{av}} N_{av}^n}{n!}, \tag{10}$$

where $N_{av} = \pi R^2 \lambda_{ap} P_{av}$ is the average number of available APs.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMC.2018.2815702, IEEE Transactions on Mobile Computing

3.4 Queueing Model of Data

To describe the dynamics of the queue in a typical IMM, we define the data queue $U(t)$ as the untransmitted data in the typical IMM at each time slot t. The queue $U(t)$ evolves according to

$$U(t+1) = \max[U(t) - u(t), 0] + a(t), \tag{11}$$

where $u(t)$ and $a(t)$ is the number of successfully transmitted and the arriving packets at the time slot t. The dynamic update rule in (11) satisfies following lemma which is useful while we analyze the performance of our proposed network association.

Lemma 1. For positive real numbers X, Y, μ, υ satisfying

$$Y = \max[X - \mu, 0] + \upsilon,$$

then the following inequality holds [33]

$$Y^2 \leq X^2 + \mu^2 + \upsilon^2 - 2X(\mu - \upsilon). \tag{12}$$

At each time slot t, the service rate of the IMM is determined by the number of accessed APs. The decision space of the IMM is denoted as $\mathbb{D}_t$:

$$\mathbb{D}_t = \begin{cases} \{1, \ldots, N(t)\}, \text{if } N(t) \neq 0 \\ \{0\}, \text{if } N(t) = 0, \end{cases} \tag{13}$$

where $N(t)$ is the total available accessible AP at time slot t and $n(t) \in \mathbb{D}_t$ is denoted as the number of connected APs at the time slot t.

Because an IMM can connect to one or multiple APs in the same channel simultaneously, the capacity of the transmission link can be regarded as a single-input-multiple-output (SIMO) channel. The capacity of SIMO with $n(t)$ accessed APs can be expressed as $\log(1 + \sum_{i=1}^{n(t)} SIR_i)$. It means that the summation of SIRs at different links determines whether the outage happens or not. Therefore, the $p_n(t)$ non-outage probability with $n(t)$ APs at time slot t can be expressed as

$$\begin{aligned} p_n(t) &= \mathbb{P}(\sum_{i=1}^{n(t)} SIR_i \geq \theta) \\ &\cong 1 - \mathbb{P}(SIR_i < \theta)^{n(t)} \\ &= 1 - (1-p)^{n(t)}, \end{aligned} \tag{14}$$

Because only one packet is transmitted in each time slot, then the $u(t)$ can be expressed as

$$u(t) = \begin{cases} 1, \text{if } \sum_{i=1}^{n(t)} SIR_i \geq \theta \\ 0, \text{if } \sum_{i=1}^{n(t)} < \theta. \end{cases}$$

Combing with (14), the mean number of serviced packets at time t condition with a decision $n(t)$ is

$$\mathbb{E}(u(t)|n(t)) = 1 - (1-p)^{n(t)}. \tag{15}$$

3.5 Minimal Required APs Network Density

Because the time-average service rate of an IMM is determined by its available APs, or the density of APs, it is necessary to know the minimal density of the APs which are able to support the fundamental horizontal connection. In the following, we denote the time-average service rate $u(t)$ and arrival packets $a(t)$ as

$$\begin{aligned} \overline{u} &\triangleq \lim_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} u(t) \\ \overline{a} &\triangleq \lim_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} a(t) \end{aligned} \tag{16}$$

Given $\overline{a}$, the minimal required number of APs without the assistance of vertical connection can be described by the following theorem.

Theorem 1. Without the assistance of vertical connection, stability of the data queue $U(t)$ cannot be guaranteed if

$$\lambda_{ap} < \log\left(\frac{1}{1-\overline{a}}\right) \Big/ (\pi R^2 P_{av} p) \tag{17}$$

Proof: To maximize the mean service rate of the typical IMM, the most intuitive way is to allow the typical IMM to utilize all APs in the transmission region and other IMMs still utilize one AP. If the service rate with this strategy still cannot support the packet arrival rate $\overline{a}$ of the IMM, then the execution of vertical connection is necessary to guarantee the data $U(t)$ to be stable. We denote the number of successfully serviced packets with fully-utilizing-AP strategy at time t as $u_{max}(t)$. Its average value is

$$\begin{aligned} \mathbb{E}(u_{max}(t)) &= \sum_{n=0}^{\infty} (1 - (1-p)^n) \frac{e^{-N_{av}} N_{av}^n}{n!} \\ &= 1 - e^{-N_{av}}, \end{aligned} \tag{18}$$

where $N_{av} = \pi R^2 \lambda_{ap} P_{av}$. In the following, we denote $\mathbb{E}(u_{max}(t))$ as $\overline{u}_{max}$.

According to queueing theory, the stability of queues is not guaranteed if the expectation of service rate is lower than that of arrival rate. Thus, the queue is not stable if the following equation holds.

$$\begin{aligned} \lim_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} \mathbb{E}[u(t)] &\leq \lim_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} \mathbb{E}[u_{max}(t)] \\ &< \overline{a} \end{aligned} \tag{19}$$

Substitute (18) into (19) and arrange it, then the proof finishes. □

4 Scenario With Horizontal Association Only

4.1 Problem Formulation

Because a HPN may not always have remaining channels for IMMs and IMMs may also move into the region where is out of coverage of a HPN like suburban area, it is necessary to explore efficient utilization of the limited number of APs without the assistance of vertical association. To support as more IMMs as possible, the horizontal association should be designed to minimize the number of utilized APs in average. The first reason is to reduce the computation complexity in BBUs pool. Though the combination of a large number of antennas results in a better multiplexing gain, it needs to run complicated algorithm

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMC.2018.2815702, IEEE Transactions on Mobile Computing

like successive interference cancellation to decode the signals. Second, more connected APs may increase the backhual load. In an uplink scenario, all the received signal must be forwarded to the combination point like BBUs pool to decode the signals. To decrease the burden on backhual networks, the desired number of connected APs should be just enough. On the other hand, as mentioned in the Section 3.1, an association scheme involves more control signal exchanges between not only IMMs and APs but also core networks. Also, decreasing the utilization number of APs can help to leave more available APs for other IMMs which may suffer from some emergency like car crash.

To overcome these challenges, we formulate the horizontal network association problem without the vertical connection as the following dynamic optimization problem.

Horizontal Network Association Problem:

$$\begin{aligned} &\min_{n(t)\in\mathbb{D}(t)} \lim_{T\to\infty} \frac{1}{T}\sum_{t=1}^{T} n(t) \\ &\text{subject to } \lim_{T\to\infty} \frac{1}{T}\sum_{t=1}^{T} \mathbb{E}[U(t)] < \infty, \end{aligned} \tag{20}$$

In this dynamic optimization problem, we try to minimize the time-average number of utilized APs. From queueing theory, a queue is stable if and only if time-average service rate is larger than that of arrival rate. Therefore, this constraint is equivalent to $\overline{u} > \overline{a}$.

4.2 Proposed Horizontal Network Association

To solve the dynamic optimization problem, a common way is to model the problem as a MDP and apply the policy or value iteration algorithm [34]. However, it may take a lot of time before the decision policy becomes stationary. Thus, this approach may not be appropriate for the mobile networks, which has highly requirement on the delay performance. Therefore, we present the following horizontal scheme based on the *Laypunov optimization*.

Horizontal Network Association Scheme: At every time slot, given the observation of the $U(t)$, the number of APs $n(t)$ follows

$$n(t) = \begin{cases} 1, \text{if } \left\lceil \frac{\ln \frac{V}{2U(t)p}}{\ln(1-p)} \right\rceil < 1 \\ \left\lceil \frac{\ln \frac{V}{2U(t)p}}{\ln(1-p)} \right\rceil, \text{if } 1 \le \left\lceil \frac{\ln \frac{V}{(2U(t)p)}}{\ln(1-p)} \right\rceil \le N(t) \\ N(t), \text{if } \left\lceil \frac{\ln \frac{V}{2U(t)p}}{\ln(1-p)} \right\rceil > N(t) \end{cases} \tag{21}$$

where $\lceil . \rceil$ represents a ceiling function and $V > 0$ is a control parameter to weight the importance of the utilization of the APs. Here we need to note that $n(t) = 0$ if the number of available APs $N(t) = 0$.

Theorem 2. With horizontal network association scheme in (21), the upper bound of the queue size satisfies

$$\lim_{T\to\infty} \frac{1}{T}\sum_{t=1}^{T} \mathbb{E}[U(t)] \le \frac{B_{max} + V\lambda_{ap}\pi R^2 P_{av}}{2\epsilon}, \tag{22}$$

where $B_{max} = \mathbb{E}(u^2_{max}(t) + a^2(t))$ and $\epsilon > 0$ satisfies $\mathbb{E}(u(t)) - \mathbb{E}(a(t)) > \epsilon$. The time-average utilization number of APs approaches the optimal utilization number n^* and satisfies

$$\lim_{T\to\infty} \frac{1}{T}\sum_{t=1}^{T} \mathbb{E}[n(t)] \le n^* + \frac{B_{max}}{V}. \tag{23}$$

Proof: Here, we define the *Lyapunov function* as $L(U(t)) = U^2(t)$ and the one-step conditional *Lyapunov drift function* $\triangle L(t)$ is defined as

$$\triangle L(t) \triangleq \mathbb{E}[L(t+1) - L(t) | U(t)]. \tag{24}$$

According to *Lemma* 1, the following inequality always holds

$$U^2(t+1) \le U^2(t) + u^2(t) + a^2(t) - 2U(t)[u(t) - a(t)]. \tag{25}$$

Taking expectation of the equation above with respect to $U(t)$, then $\mathbb{E}(\triangle U(t))$ satisfies

$$\begin{aligned} \mathbb{E}[\triangle L(t)] &\le \mathbb{E}\left[u^2(t) + a^2(t)\right] - 2\mathbb{E}[U(t)(u(t) - a(t))] \\ &\le \mathbb{E}\left[u^2_{max}(t) + a^2(t)\right] - 2\mathbb{E}[U(t)(u(t) - a(t))] \\ &= B_{max} - 2\mathbb{E}[U(t)(u(t) - a(t))] \end{aligned} \tag{26}$$

Adding $V\mathbb{E}(n(t))$ to both side of (26) and arrange it, we can get

$$\begin{aligned} &\mathbb{E}[\triangle L(t) + Vn(t)] \le \\ &B_{max} + 2\mathbb{E}[U(t)a(t)] - \mathbb{E}[2U(t)u(t) - Vn(t)]. \end{aligned} \tag{27}$$

The horizontal network association actually tries to minimize the right hand side of the equation above given the $U(t)$. That is,

$$\max_{n(t)\in\mathbb{D}(t)} \quad \mathbb{E}[2U(t)(u(t) - a(t)) - Vn(t) | U(t)]. \tag{28}$$

Substitute (15) into (28) and choose the integer to maximize the equation above, we get the *horizontal network association* in (21). The details of the proof are shown in Appendix A.

We denote the optimal average utilization number of APs as n^* for the problem (20). Regarding the $n(t)$ as the resulting random process of the proposed scheme, then we can find that

$$\begin{aligned} &\mathbb{E}[\triangle L(t) + Vn(t)] \\ &\le B_{max} - 2\mathbb{E}[u(t) - a(t)) + V\mathbb{E}(n(t)] \\ &\le B_{max} - \mathbb{E}[2U(t)(u(t) - a(t))] + Vn^*, \end{aligned} \tag{29}$$

the second inequality comes from the fact of minimization in (28).

Taking the time average from $t = 1 \to \infty$ and rearrange the equation above, we can get the follows

$$\begin{aligned} &\lim_{t\to\infty} \frac{1}{T}\sum_{t=1}^{T} \mathbb{E}[U(t)] \\ &\le \frac{B_{max} + Vn^*}{2\epsilon} - \lim_{T\to\infty}\left(\frac{\mathbb{E}(U^2(T+1) - U^2(1))}{T}\right) \\ &\le \frac{B_{max} + V\mathbb{E}\left(\lim_{T\to\infty} \frac{1}{T}\sum_{t=1}^{T} N(t)\right)}{2\epsilon} \\ &= \frac{B_{max} + V\lambda_{ap}\pi R^2 P_{av}}{2\epsilon}, \end{aligned}$$

which is the upper bound of the average size of the data queue $U(t)$. The second inequality comes from the fact that the IMMs can at most access the all APs in the decision space $\mathbb{D}(t)$.

To prove the inequality (23), we move $\triangle U(t)$ in (29) to the right hand side and take the time average from $t \to \infty$

$$\begin{aligned} &\lim_{T\to\infty} \frac{1}{T}\sum_{t=1}^{T} \mathbb{E}[n(t)]V \\ &\le B_{max} + Vn^* - \lim_{T\to\infty} \frac{\mathbb{E}\left[U^2(T+1) - U^2(1)\right]}{T} \\ &\le B_{max} + Vn^*, \end{aligned} \tag{30}$$

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMC.2018.2815702, IEEE Transactions on Mobile Computing

the last inequality comes from the fact that $\lim_{T\to\infty}\frac{\mathbb{E}(U^2(T+1)-U^2(1))}{T}$ approaches to 0 if T goes to infinity. By dividing both sides of (30) with V, we get the inequality (23). □

From *Theorem* 2, we know that there is a tradeoff between the utilization number of the available APs and the resulting queue length. We can interpret V as the cost of utilizing the APs. By decreasing V, the size of the queue decreases but the utilization number of the APs increases. On the contrary, we can also decrease the utilization of the APs by increasing V, but it results in longer delay.

5 Scenario With Horizontal and Vertical Association

5.1 Problem Formulation

Due to highly mobility of IMMs, the stability condition in (17) may not always hold. For example, a vehicle gets into the downtown or it is in the rush hours, the volume of vehicles flows increases significantly and the stability condition cannot be guaranteed. In such case, a possible solution is to build a *dual connectivity*, by which, IMMs can connect to APs and a HPN simultaneously. *Dual connectivity* can offload some of data flows into the HPN network to reduce the burden of the AP networks. The idea of utilizing HPNs to relieve the burden of data traffic can be traced back to [35]. However, the procedure of building a vertical connection involves lots of information exchanges, *e.g.*, authentication, between IMMs and core networks as mentioned in Section 3.1. To avoid the additional burden on core networks and the complexity of vertical connections, IMMs should not aggressively execute vertical network association as possible as they can.

We denote $a(t)$ as the total arrival data at time t and $a_A(t) \le a(t)$ as the data transmitted through AP networks at time t. The amount of $a(t)-a_A(t)$ is offloaded to a HPN network through the vertical connection. The update rule of the data queue $U(t)$ is

$$U(t+1) = \max[U(t)-u(t),0]+a_A(t). \tag{31}$$

To minimize the average amount of data flowing through the vertical connection, *i.e.*, the HPN network, it is equivalent to maximize the time average of $a_A(t)$. Therefore, we formulate a vertical network association problem as follow. *Vertical Network Assocation Problem:*

$$\begin{aligned} \max_{n(t)\in\mathbb{D}(t), a_A(t)\le a(t)} \quad & \lim_{T\to\infty}\frac{1}{T}\sum_{t=1}^{T} a_A(t) \\ \text{subject to} \quad & \overline{n} \le N_{av} \\ & \lim_{T\to\infty}\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}[U(t)] \le \infty, \end{aligned} \tag{32}$$

where $\overline{n} = \lim_{T\to\infty} 1/T\sum_{t=1}^{T}\mathbb{E}(n(t))$ is the time-average utilization number of the APs and N_{av} is the average available APs. The first constraint in (32) comes the fact that the average utilization number of the APs should not exceed the average available APs. The second one is to guarantee the stability of the queue.

5.2 Virtual Utilization Queue

To tackle the constraints in the dynamic programming problem in (32), a novel approach is to build the *virtual queue* [36]. A *virtual queue* is like a budget table. At each beginning of a time slot, an IMM get a quota N_{av} on the utilization number of APs. The IMM records the difference between utilization and the quota ($n(t)-N_{av}$). This budget table is known as *virtual queue*. We denote this *virtual queue* as $X(t)$ and its update rule is

$$X(t+1) = \max[X(t)-N_{av},0]+n(t), \tag{33}$$

where N_{av} is the average number of available APs in the environment. From the queueing theory, the stability of the queue $X(t)$ is

$$\overline{n} \le N_{av}, \tag{34}$$

which is the constraint in (32). In this way, a time dynamic constraint problem can be converted into the stability problem and all the *Lyapunov optimization* can be applied.

5.3 Proposed Horizontal and Vertical Network Association

To improve end-to-end delay performances, reduction of control signal plays an important role [10]. On the other hand, due to the existence of a large amount of IMMs in a network, a significant portion of spectrum may be occupied by the control signal exchanges if all the IMMs ask for the service of HPN and thus harm the delay performance. Therefore, vertical association, which introduces additional control signals between the IMMs and the networks, should be triggered only if the delay requirement of the data queue cannot be supported. To achieve this goal, we proposed a delay-aware vertical association based on the *Lyapunov optimization*, which ensures seamless connection for the IMMs. To decrease the burden of the core network or BBUs, the vertical association should be triggered by the IMMs instead of the network side.

Vertical Network Association: At every time slot, given the observation of $U(t)$ and $X(t)$, the number of accessing APs $n(t)$ follows

$$n(t) = \begin{cases} 1, \text{if } \left\lceil \frac{\ln\frac{X(t)}{U(t)p}}{\ln(1-p)} \right\rceil < 1. \\ \left\lceil \frac{\ln\frac{X(t)}{U(t)p}}{\ln(1-p)} \right\rceil, \text{if } 1 \le \left\lceil \frac{\ln\frac{X(t)}{U(t)p}}{\ln(1-p)} \right\rceil \le N(t) \\ N(t), \text{if } \left\lceil \frac{\ln\frac{X(t)}{U(t)p}}{\ln(1-p)} \right\rceil > N(t)) \end{cases} \tag{35}$$

The new arriving data are transmitted via the vertical association whenever $U(t) > V/2$, or else the arriving data are transmitted through APs networks. Here we need to note that $n(t) = 0$ if the number of available APs $N(t) = 0$.

The proposed vertical association scheme has the properties described in *Theorem* 3.

Theorem 3. We denote the ratio of the datas transmitted via vertical connection as P_v. The upper bound of P_v is

$$P_v \triangleq \frac{\overline{a}-\overline{a}_A}{\overline{a}} \le 1 - \frac{\overline{u}_{max}-(B_{max}+C_{max})/V}{\overline{a}}, \tag{36}$$

where $\overline{a}_A = \lim_{t\to\infty}\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}(a_A(t))$, B_{max} and C_{max} are

$$\begin{aligned} B_{max} &= \mathbb{E}\left(u^2(t)+a^2(t)\right) \\ C_{max} &= \mathbb{E}\left(N_{av}^2+n^2(t)\right) \end{aligned} \tag{37}$$

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMC.2018.2815702, IEEE Transactions on Mobile Computing

The average size of queue $\mathbb{E}(U(t))$ is upper bounded by $V/2+\mathbb{E}(a(t))$.

Proof: The *Lyapunov function* is defined as $L(t)=U^2(t)+X^2(t)$ and the *Lyapunov drift* function is

$$\triangle L(t) \triangleq \mathbb{E}[L(t+1)-L(t)|U(t),X(t)]. \tag{38}$$

According to *Lemma* 1, the following inequality holds

$$\begin{aligned}\mathbb{E}(\triangle U^2) &= \mathbb{E}\left[U^2(t+1)-U^2(t)\right]\\ &\leq B_{max}-2\mathbb{E}[U(t)u(t)-a_A(t)] \qquad (39)\\ \mathbb{E}(\triangle X^2) &= \mathbb{E}\left[X^2(t+1)-X^2(t)\right]\\ &\leq \mathbb{E}\left[N_{av}^2+n_{max}^2(t)\right]-2\mathbb{E}[X(t)(N_{av}-n(t))]\\ &= C_{max}-2\mathbb{E}[X(t)(N_{av}-n(t))], \qquad (40)\end{aligned}$$

where $n_{max}(t)$ is the utilization number of APs with the fully-utilizing strategy $n(t)=\max(\mathbb{D}(t))$. Combining (39) and (40) then adding $-V\mathbb{E}(a_A(t))$ at both side, we get

$$\begin{aligned}&\mathbb{E}[\triangle L(t)]-V\mathbb{E}[a_A(t)] \leq B_{max}+C_{max}-2\mathbb{E}[X(t)N_{av}]\\ &-2\mathbb{E}[U(t)u(t)-X(t)n(t)]+\mathbb{E}[2U(t)a_A(t)-V\mathbb{E}(a_A(t))]\end{aligned} \tag{41}$$

The proposed vertical association scheme is actually to minimize the right hand side of (41). Given $X(t)$ and $U(t)$, we solve the following two optimization problems.

$$\max_{n(t)\in\mathbb{D}(t)} \quad \mathbb{E}[U(t)u(t)-X(t)n(t)|U(t),X(t)] \tag{42}$$

$$\min_{a_A(t)\leq a(t)} \quad \mathbb{E}[2U(t)a_A(t)-Va_A(t)|U(t)]. \tag{43}$$

To maximize (43), we just need to check whether $2U(t)-V$ is larger than 0 or not. If $2U(t)-V>0$, then $a_A(t)=a(t)$. Otherwise, the IMM executes vertical connection to offload $a(t)$ to HPNs (then $a_A(t)=0$). To get the results in (35), we can follow the similar procedures of the proof in Appendix A to maximize (42).

With the proposed vertical association scheme, the following inequality holds.

$$\begin{aligned}&\mathbb{E}[\triangle L(t)]-V\mathbb{E}[a+A(t)] \leq B_{max}+C_{max}\\ &-2\mathbb{E}[2X(t)(N_{av}-n(t))]-\mathbb{E}[2U(t)(u(t)-a_A(t))]-Va^o,\end{aligned} \tag{44}$$

where a^o is the average data traffic flowing through APs network with any other arbitrary horizontal association scheme. The inequality comes from minimization of right hand side of (41). By taking the time average on both side in (41), we get

$$\begin{aligned}\lim_{T\to\infty}\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}[a_A(t)] &\geq a^o-\frac{B_{max}+C_{max}}{V}\\ &= \overline{u}_{max}-\frac{B_{max}+C_{max}}{V}.\end{aligned} \tag{45}$$

Equality in (45) holds for every a^o with any arbitrary scheme. To minimize the gap between left hand and right hand side, we set $a^o=\overline{u}_{max}$. The $\mathbb{E}(u_{max}(t))$ is the average service rate with fully-utilizing APs as shown in (18). Due to $P_v \triangleq \frac{\overline{a}-\overline{a}_A}{\overline{a}}$, we find

$$\begin{aligned}P_v &\triangleq \frac{\overline{a}-\overline{a_A}}{\overline{a}}\\ &\leq 1-\frac{\overline{u}_{max}-(B_{max}+C_{max})/V}{\overline{a}}.\end{aligned} \tag{46}$$

The upper bound of expectation of data queue $\mathbb{E}(U(t))$ comes from the fact that the queue stops to access newly data and offloads these data to the vertical connection if $U(t)>V/2$. Therefore, the maximal size of the data queue is $V/2+a(t)$. Takes the average of it and the proof finishes.

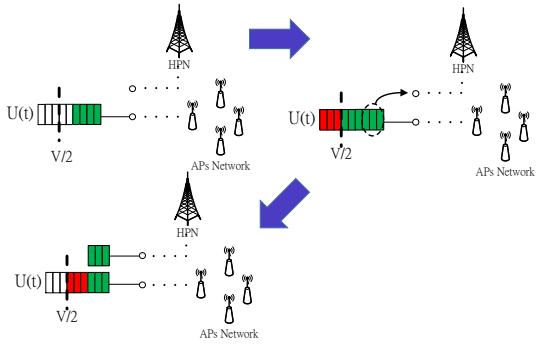

Fig. 4: The traffic flows are switched to the HPN once the data queue $U(t)$ is not larger than $V/2$.

□

The illustration of the vertical association scheme is shown in Fig. 4. To follow First-Come-First-Service (FCFS) principle, the packets exceeding $V/2$ are not directly switched into HPNs. Instead, the first few packets are switched to the vertical connection until the total queue $U(t)<V/2$.

6 Delay Violation Probability

Effective bandwidth and *effective capacity* [37], [38] are shown useful to comprehend QoS in a time-varying wireless channel. Recent years, the theories are also utilized to the QoS performance of power control [39], radio resource allocation [40] and computational management [41]. In this section, we explore the QoS stability of the proposed scheme from the viewpoint of *effective bandwidth* and *effective capacity*.

For user experience, we may not only care about the expectation of a queue size but also the probability of a queue size growing larger than a threshold B. This is called the delay violation probability P_{vi}, *i.e.*,

$$P_{vi} \triangleq \lim_{t\to\infty}\mathbb{P}(U(t)\geq B). \tag{47}$$

To design a system which can guarantee that P_{vi} is smaller than some required probability, *effective bandwidth capacity theory* is an useful tool [42]. *Effective bandwidth* specifies the minimal constant service rate c that can support a given arriving data stream to satisfy the required delay violation probability. *Effective bandwidth* is defined as

$$E_B(s) \triangleq \lim_{t\to\infty}\frac{1}{st}\ln\mathbb{E}\left(e^{sA(t)}\right), \tag{48}$$

where $s>0$ and $A(t)$ is an accumulated data arrival process, *i.e.*, $A(t)=\sum_{k=0}^{t}a_A(k)$.

The concept of *effective capacity*, which is the duality of *effective bandwidth*, is defined as

$$E_C(s) \triangleq \lim_{t\to\infty}-\frac{1}{st}\ln\mathbb{E}\left(e^{-sS(t)}\right), \tag{49}$$

where $S(t)=\sum_{k=1}^{t}u(k)$ is an accumulated serviced data from beginning to time slot t. *Effective capacity* specifies the maximal constant data arrival rate that the system can support such that the required delay violation probability can be satisfied.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMC.2018.2815702, IEEE Transactions on Mobile Computing

Because a queue size at time $U(t)$ can be expressed as $U(t) = A(t) - S(t)$, thus the delay violation probability P_{vi} can be further expressed as

$$\begin{aligned} P_{vi} &\triangleq \lim_{t\to\infty} \mathbb{P}(U(t) \geq B) \\ &= \lim_{t\to\infty} \mathbb{P}(A(t) - S(t) \geq B) \\ &\leq \frac{\mathbb{E}\left(e^{sA(t)}\right)\mathbb{E}\left(e^{-sS(t)}\right)}{e^{sB}}. \end{aligned} \tag{50}$$

We can take logarithm on the both sides and get

$$\begin{aligned} \ln P_{vi} &\leq \lim_{t\to\infty} \ln\mathbb{E}\left(e^{sA(t)}\right) + \ln\mathbb{E}\left(e^{-sA(t)}\right) - sB \\ &= \lim_{t\to\infty} st\left(\frac{1}{st}\ln\mathbb{E}\left(e^{sA(t)}\right) - \frac{-1}{st}\ln\mathbb{E}\left(e^{-sS(t)}\right)\right) - sB \\ &= \lim_{t\to\infty} st\left(E_B(s) - E_C(s)\right) - sB. \end{aligned} \tag{51}$$

Above equation is meaningful only if there exists a $s^* > 0$ such that

$$E_B(s^*) = E_C(s^*). \tag{52}$$

Before finding the solution existence condition of (52), we first need to know that the mean of data arrival rate can be shown to be $\overline{a} = \lim_{s\to 0} E_B(s)$ and the mean of data service rate is $\overline{u} = \lim_{s\to 0} E_C(s)$. Second, *effective bandwidth* $E_B(s)$ is an increasing function, thus, *effective capacity* is a decreasing function [38]. Therefore, the solution exists only if $\overline{a} < \overline{u}$, which is also the condition that the solutions of the proposed problems in (20) and (32) exist. If the solution of (52) exists, we can further get

$$P_{vi} \leq e^{-s^*B}, \tag{53}$$

where s^* is a constant such that $E_B(s^*) = E_C(s^*)$. Therefore, we can conclude the following theorem.

Corollary 1. If the stable condition in *Theorem* 1 is satisfied, then the upper bound of the delay violation probability of the proposed scheme follows exponential decay function, even without the assistance of the vertical network association.

7 Design and Complexity

7.1 Design Procedure

Generally speaking, a network association scheme in heterogeneous networks consists of three different phases. (1) discovery of newly encountered APs (2) decision on network connection (3) execution of the connection. In the first phase, APs should periodically advertise the control signals such as reference signals to inform IMMs the existence of the APs. In the second phase, the decision about connecting to APs (horizontal connection) or to a HPN (vertical connection) is made. In our proposed scheme, each IMM proactively triggers such decision procedure based on own information (size of their data queue) and discovery of the APs. In the third phase, the data packets are routed to newly connected APs or a HPN if the vertical connection is needed. This phase includes the authentication, authorization, *etc*. Because the procedure of establishing a vertical connection involves additional control signal exchanges, we maximize the data traffics flowing through the APs in (32).

Illustrated as Fig. 5, the proposed proactive network association consists of the following procedures.

1) IMMs scan the control channels to find total available APs in the dedicated channel.

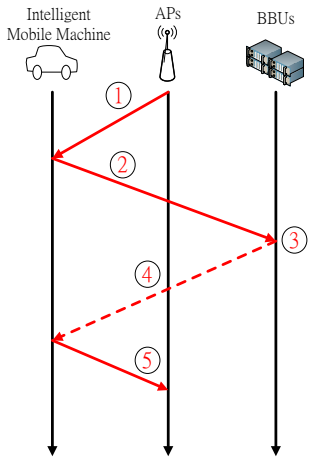

Fig. 5: Time step diagram for the *horizonal* and *vertical network association*.

2) IMMs inform the BBUs pool via one or multiple discovered APs about the number of needed APs according to (21) (without the vertical connection scenario) or (35) (with vertical connection scenario). If building a vertical connection is necessary, IMMs also inform the BBUs pool in this phase.
3) At this moment, the BBUs pool chooses a set of APs through which the information comes from the IMMs and decode them. All the received signals are processed by the BBUs pool, therefore, there is no need to inform the IMMs about the APs which BBUs pool chooses.
4) The BBUs pool informs IMMs about the available channels in a HPN if the execution of vertical connection is necessary.
5) IMMs start to synchronize with the (horizontal and vertical) channels and then transmit data.

On the other hand, to obtain the accurate value of non-outage probability p in (5), a HPN can periodically broadcast this information to all IMMs or via distributed density estimation [43] among IMMs.

7.2 Complexity Discussion

We discuss the algorithm complexity in this section. Our proposed algorithm adopts a vehicle-centric approach to determine the "enough" number of the connected APs to avoid the complexity coming from CoMP.

In the scenario of only horizontal network association available, an IMM determines the number of connected APs by (21). All the necessary information like $U(t)$ can be obtained from its own information and environmental setting V and p. Similarly, in the scenario with vertical network association, an IMM determines whether or not to connect with HPN solely depends on its queueing size being larger than $V/2$ or not as shown in (35). In the procedure of obtaining the number of connected APs, an IMM only needs to observe its own queueing size (including the virtual queue $X(t)$) and the environment setting V and p.

The primary complexity comes from coordinating among IMMs, APs and HPNs. It is no doubt that integrating all the infrastructure resources to form a CoMP network is able to improve the overall performance. However, this approach also comes

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMC.2018.2815702, IEEE Transactions on Mobile Computing

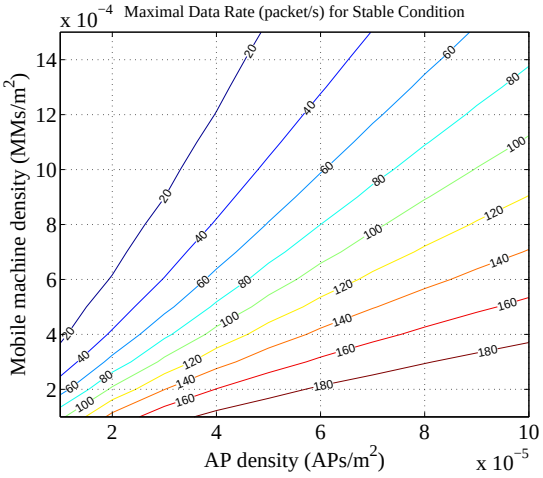

Fig. 6: Maximal data rate to guarantee the stable condition in (17) without the assistance of vertical association.

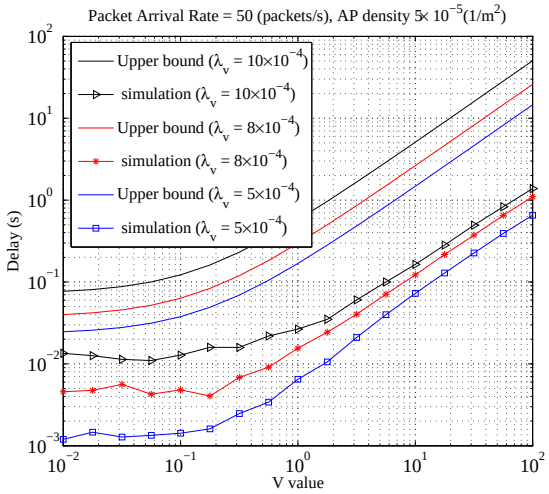

Fig. 7: Larger V can be interpreted as higher cost of utilizing APs. Thus, it results in a less service rate and a longer waiting time in queue (delay).

with increased complexity due to additional authorization procedures, pilots signals, synchronization issues and signal processing. Complexity increases with the number of APs [22]. Take this complexity into consideration, a load-aware approach to control the size of the CoMP also attracts other research's interests [25]. Based on the IMMs' available information, such vehicle-centric approach does not rely on frequent control signalling and thus obtain better delay performance. To strike the balance between the complexity and the performance gain from CoMP, we propose the algorithm to satisfy the delay requirements and simultaneously keep the utilization number of APs be as small as possible.

8 Performance of Network Association Scheme

8.1 Simulation Result

In the simulation, we use the parameters setting as $R = 200$, $v = 15$, $J = 20$, $\theta = 3$. Considering the Doppler shift effect, the coherence time is about $5ms$ under the velocity $v = 15m/s$ ($55km/h$) with a carrier frequency $2GHz$. Therefore, we set the duration of time slot $5ms$. We set the non-outage threshold $\theta = 3$ to guarantee at least QPSK being reliably transmitted per symbol time. Due to obstruction effect of the buildings, we set the path-loss exponent $\alpha = 6$. The total duration time of the simulation is $1000s$ and the iteration times is 500. In the vertical association scheme simulation, we assume that each IMM can be allocated an independent channel from HPN thus each packet switched to vertical association can be successfully transmitted in one time slot without interference. To avoid different processing time of the hardwares, the delay performance refers to the waiting time in queues. The additional delay caused by the coordination of the multiple APs and the HPN in the backhaul networks are captured by the utilization rate.

To guarantee the stability of the queues under the scenario without the assistance of the vertical network association, the most important thing is to ensure that the proposed scheme operates under the solution existing condition (*i.e.*, *Theorem* 1). Fig. 6 illustrates the contour plot of the maximal packet arrival rate corresponding to the different densities of APs. The numeric value on the line represents the maximal allowable data arrival rate. The stability conditions can be guaranteed only if the operating point

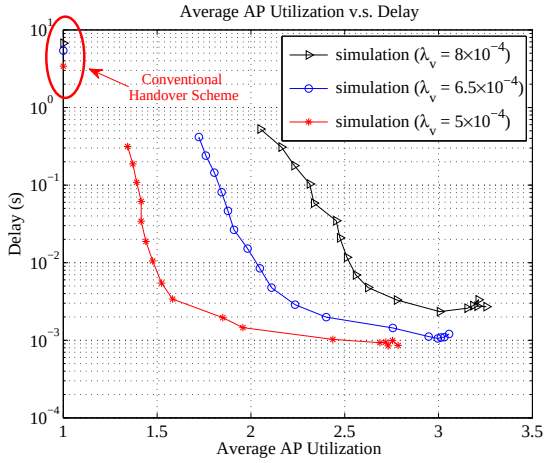

Fig. 8: The proposed scheme can dynamically control the number of APs and thus can achieve a better delay performance.

(density of APs and IMMs) is on the left-upper side of the line. We can find that the bottom right of the figure can support faster data transmission. It makes sense that each IMM can transmit faster given more resource (available APs) with less competitors.

In Fig. 7, we illustrate the queue delay without the vertical association scheme corresponding different V values. The time-average packet arrival rate is $\overline{a} = 50$(packets/s). The engineering meaning of V can be interpreted as the cost of utilizing an AP. With the larger V, the fewer APs are utilized and results in a slower service rate and longer waiting in a queue. We find that the delay performance can be bounded by our analysis results in *Theorem* 2. By adjusting V, we get the desired delay performance solely by the analytical results. In Fig. 8, we compared our scheme with the conventional handover scheme. The conventional handover scheme refers to that an IMM connects to only one AP. To obtain the best SIR, an IMM always connects to an AP with the shortest distance. We find that the proposed scheme outperforms the conventional handover scheme in terms of the delay performance. Because the conventional scheme, which utilizes only one AP,

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMC.2018.2815702, IEEE Transactions on Mobile Computing

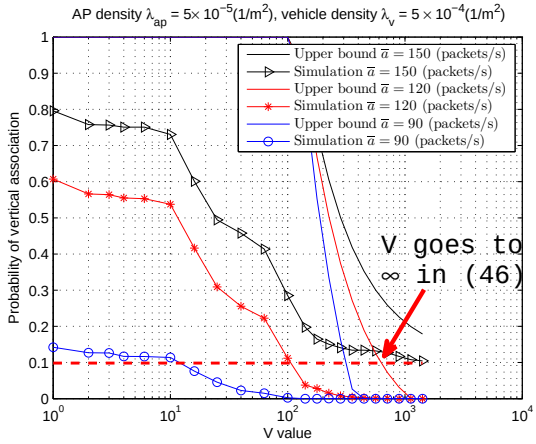

Fig. 9: The packet arrival rate exceeds the maximal allowable arrival rate, building a vertical association becomes necessary even if V goes to infinity.

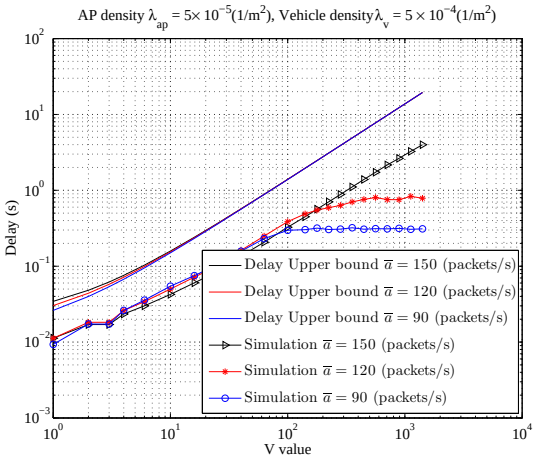

Fig. 10: If V is small, the IMMs with larger average packet arrival rate $\overline{a}$ can slight benefit from frequent utilization of the vertical association.

lacking flexibility, the queue cannot be alleviated successfully. However, if we consider the proposed multiple-to-multiple network association, the delay performance can be improved largely. It should be noted that the proposed scheme can dynamically adjust the number of the connected APs. It connects more APs only if the size of queues keep growing. Therefore, the proposed scheme also achieves AP-utilization efficiency.

In Fig. 9, we illustrate the probability of executing the proposed vertical association P_v with different $\overline{a}$ and V. According to the figure, we can find that P_v is smaller if the value of V is larger. However, if the packet arrival rate $\overline{a}$ exceeds the maximal allowable rate of the stable condition in (17), the execution of vertical association becomes necessary no matter how large V is. By limiting $V \to \infty$ in (46) and $P_v = 1 - \frac{\overline{a}_{max}}{\overline{a}}$, which means that the ratio of data shall be directed to the vertical association if an IMM utilizes all the available horizontal APs.

Fig. 10 shows the average delay corresponding to different packet arrival rates. It shows that all the delays are upper bounded by the bound provided in *Theorem* 3. In the figure, an interesting phenomenon is that the data with larger packet arrival rate experi-

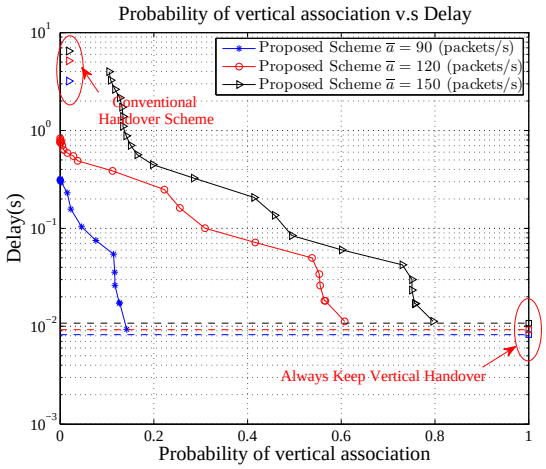

Fig. 11: Due to the integration of the distributed APs and HPN, the proposed scheme can reach the same delay performance as that of always utilizing HPN.

ences a smaller delay if value V is small. This benefit comes from that the data with larger packet arrival rate utilizes more vertical resources if V is small, as shown in Fig. 9. If V increases, IMMs tend to utilize more horizontal association. The average delay with the smallest packet arrival rate first reaches stability, even without the assistance of vertical association.

In Fig. 11, we compare our proposed proactive network association with the conventional handover and a scheme with always maintaining a vertical connection. In the conventional handover scheme, each IMM connects to only one AP or switches to a HPN if there is no available AP in the service region. In such way, the transmission rate for an IMM cannot support the data arrival rate and thus results in large delay performance. With our simulation setting, the probability of without APs is only 0.018, *i.e.*, probability of connecting to a HPN is 0.018. Even though there are enough APs, without the flexible network resource allocation scheme, the low delay performance still cannot be achieved. With always-keeping vertical connection approach, an IMM always keeps a vertical connection with a HPN and horizontal connections with APs, which achieves the best delay performance. However, such always-connected approach may occupy all the radio resources and not be acceptable in the practical scenario, especially when the amount of IMMs is large. By fully utilizing the available APs, the proposed proactive network association successfully compromises "best possible" scenario with less vertical associations and thus less control signal exchanges in a dynamic operating environment.

8.2 Simulation with Real Mobility Data

In the previous simulations, the mobility model of the IMMs can be regarded as that all the IMMs do not follow a certain pattern but run in an arbitrary direction. This simulation environment is more similar to the suburban area. Therefore, the spatial distribution of the available APs almost follows identical independent distribution (*i.i.d*) and the expected performance can be reached. We also care about the performance in the urban area, where the vehicles follow the specially designed city roads. In this section, we utilize the practical taxi mobility data to further verify the proposed vertical and horizontal network association scheme. The taxi mobility data are collected from the operating taxis in the

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMC.2018.2815702, IEEE Transactions on Mobile Computing

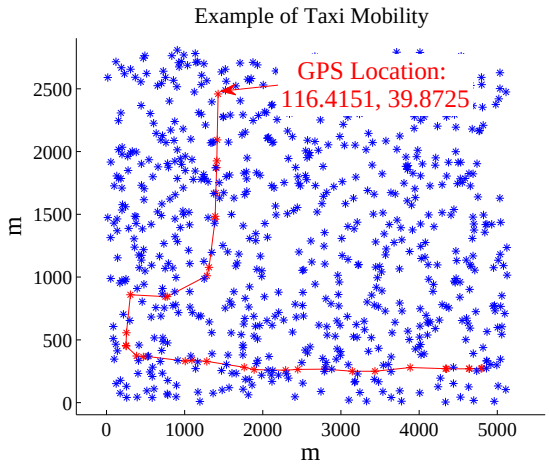

Fig. 12: Example of taxi mobility.

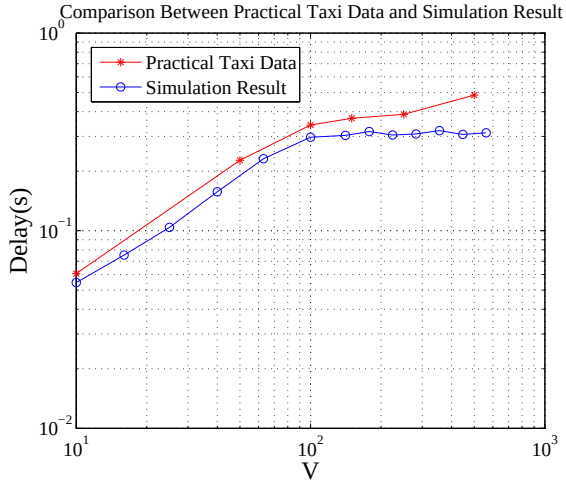

Fig. 13: The reason for the performance loss is that the taxi may go into the area lack of APs sometimes.

Beijing from 2012/11/01 to 2012/12/31. We randomly choose 100 taxis to simulate the queues variation with the proposed network association scheme and compare them with the simulation results in the previous section. Fig. 12 illustrates a randomly selected taxi. The red line is the reported location of the taxi and the blue star points are randomly distributed APs in the network with distribution density $5\times10^{-5}(1/m^2)$ and the service radius is 200(m). The GPS location of the starting point is longitude 116.4151 and latitude 39.8725. The distribution density of APs λ_{ap} is $5\times10^{-5}(1/m^2)$ and the service radius of each AP is 200(m).

Fig. 13 illustrates the delay performance with data arrival rate 90 (packets/s). Compared with the simulation environment, we can find that the taxis suffer from little performance loss especially if V is large. It is the reason that some taxis may go around an area without enough APs. We need to know that larger V means less utilization of vertical network association in the scenario with the assistance of the vertical network association. In the situation, where there is no enough AP, the vertical network association providing ubiquitous connection service plays an important role. As shown in Fig. 14, the APs cannot support all the arriving data sometimes. The size of the queues may increase rapidly until reaching the threshold of triggering vertical network association or

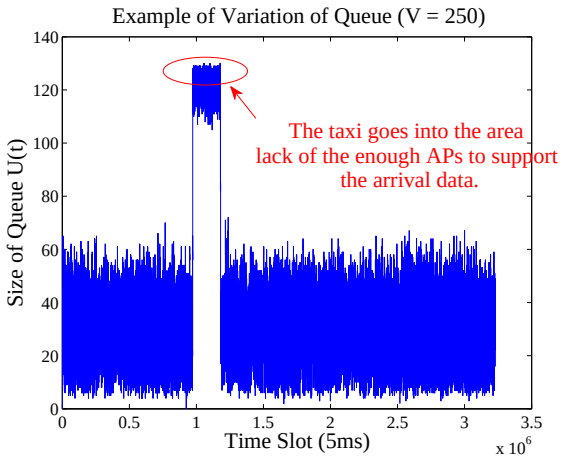

Fig. 14: The size of queue may increase fast until it triggers the vertical network association.

the taxis leave this area. If we set V large, the taxi cannot utilize the vertical network association immediately and thus results in worse delay performance. This phenomenon can be better explained with the delay violation analysis. Fig. 15 illustrates the delay violation probability under different violation thresholds without the vertical network association. Because different value of V results in different service rates of the data queues, the delay violation probability can be upper bounded by e^{-sB} with different s. However, as V increases, this upper bound may not work anymore, like the line $V = 250$. In Fig. 15, due to the taxis going around in an area without enough APs. The success of data transmissions thus highly depends on the vertical network association. For smaller V, the threshold to trigger the vertical network association is smaller, and hence the size of queue drops quickly even if the taxis are in this kind of areas. In Fig. 16, we compared with the results with the ideal movement mode. We can find that the performances are similar if the triggering threshold of vertical network association is small. If V is large ($V = 150, 250$ in Fig. 16), the simulation results are better than the one with practical taxi data. This result shows that if the movement is similar to the ideal movement, like on the highway or suburban environment, the distributed APs can support the arriving data alone. However, if the environment is similar to the urban area (as the practical data), the vertical network association plays a more important role to guarantee the delay performance. Consequently, the larger V may result in insufficient utilization of the vertical network and the delay violation probability cannot be guaranteed.

9 Conclusion

In this paper, a proactive network association scheme that can provide multiple-to-multiple switches are proposed. We regard the network association as dynamic resource allocation in heterogeneous networks, with two different types of resources: *horizontal* and *vertical associations*. This resource-allocation-based approach is quite different from conventional network association or handover technology in cellular networks. The corresponding dynamic resource allocation problems are proposed to utilize radio resources in the most efficient way. To solve the proposed dynamic optimization problem, we take the advantage of *Lyapunov optimization* to provide IMMs with insightful decision schemes to

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMC.2018.2815702, IEEE Transactions on Mobile Computing

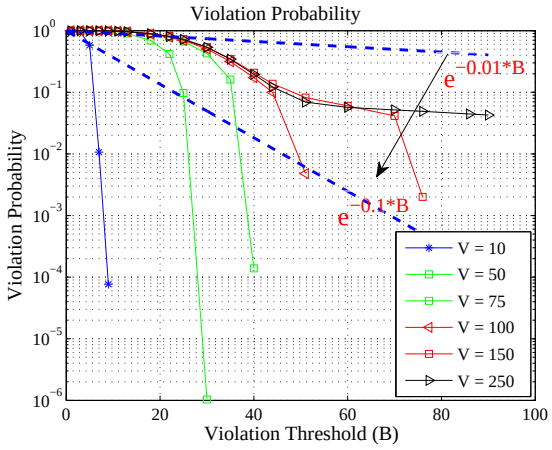

Fig. 15: Illustration of the violation probability without the vertical network association.

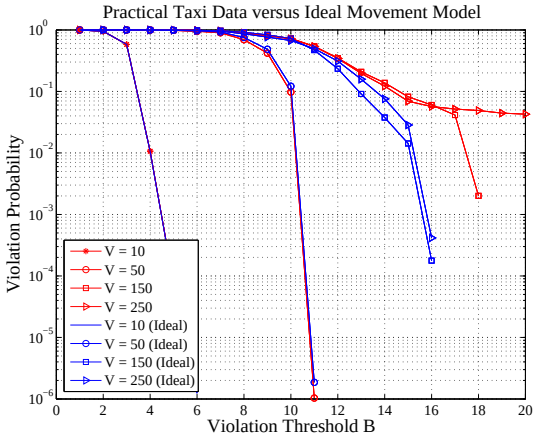

Fig. 16: Comparison of the violation probability of the ideal and practical movement data with the assistance of vertical network association.

guarantee the low-latency and ultra-reliable communication with efficacious utilization of limited distributed APs and HPNs simultaneously. The proposed proactive network association utilizes a minimal number of APs and trigger the vertical association only if it is necessary, which profits by less information exchanges and thus reduction of the delay in highly dynamic operation like vehicular networks.

Acknowledgement

This work was supported in part by the Hong Kong, Macao and Taiwan Science and Technology Cooperation Projects under Grant (2016YFE0122900 and 2014DFT10320).

Appendix A

Proof: We define $f(n(t))$ as the objective function in (28)

$$f(n(t)) = 2U(t)(1-(1-p)^{n(t)}) - Vn(t),$$

and the optimal solution as N^*. The optimal solution should satisfy two conditions. The first is

$$\begin{aligned} f(N^*) - f(N^*+1) &= 2U(t)(1-(1-p)^{N^*}) - VN^* \\ &\quad - (2U(t)(1-(1-p)^{N^*+1}) - V(N^*+1)) \\ &= -2U(t)(1-p)^{N^*}p + V \geq 0. \end{aligned}$$

We thus get

$$N^* \geq \frac{\ln \frac{V}{2U(t)p}}{\ln(1-p)} \tag{54}$$

The second condition is

$$\begin{aligned} f(N^*) - f(N^*-1) &= 2U(t)(1-(1-p)^{N^*}) - VN^* \\ &\quad - (2U(t)(1-(1-p)^{N^*-1}) - V(N^*-1)) \\ &= 2U(t)(1-p)^{N^*-1}p - V \geq 0. \end{aligned}$$

Then we get

$$N^* \leq \frac{\ln \frac{V}{2U(t)p}}{\ln(1-p)} + 1. \tag{55}$$

Combining (54) and (55), we get

$$\frac{\ln \frac{V}{2U(t)p}}{\ln(1-p)} \leq N^* \leq \frac{\ln \frac{V}{2U(t)p}}{\ln(1-p)} + 1, \tag{56}$$

(21) can thus be obtained. □

Appendix B

For each IMM enters into the transmission region with an angle θ, the service time,*i.e.*, the duration staying the circular region in Fig. 17, can be expressed as

$$\frac{2R\cos\theta}{v}.$$

Because of homogeneous property, the projection of the arrival IMMs on the vertical axis follows uniform distribution. That is, the probability of an IMM arriving with an angle θ is $\frac{R\cos\theta d\theta}{R} = \cos\theta d\theta$. Therefore, the expected duration of an IMM staying in the circular transmission region is

$$\int_0^{\frac{\pi}{2}} \frac{2R\cos\theta}{v}\frac{R\cos\theta}{R}d\theta = \frac{\pi R}{2v}.$$

To derive the expected arrival rate of the IMMs with a velocity v, we need to refer to the Fig. 17. We first calculate the "area increased" rate as shown in the shadowed area in Fig. 17. In the figure, the incremental area at the angle θ is $Rv\cos\theta dt$. To obtain the shadowed area, we integrate θ from $-\frac{\pi}{2}$ to $\frac{\pi}{2}$.

$$\int_{-\frac{\pi}{2}}^{\frac{\pi}{2}} R\cos\theta v d\theta dt = 2Rvdt. \tag{57}$$

The spatial distribution of the APs follows a PPP with density λ_{ap} but not all the APs have remaining channels for the newly arrival IMMs. We denote the probability of an AP being available as P_{av}, *i.e.*, the probability that not all J channels are occupied by the IMMs. Therefore, the effective density is $\lambda_{ap}P_{av}$ and the expected arrival rate of the IMMs is $2Rv\lambda_{ap}P_{av}$.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMC.2018.2815702, IEEE Transactions on Mobile Computing

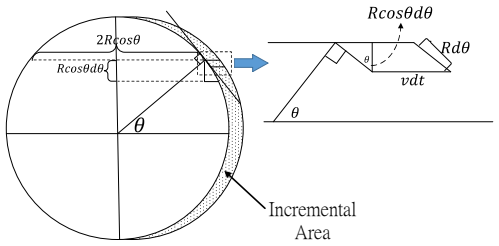

Fig. 17: Illustration of the expected time staying in the transmission region and the expected arrival rate.

REFERENCES

[1] S. L. Poczter and L. M. Jankovic, "The google car: Driving toward a better future?" *Journal of Business Case Studies (JBCS)*, vol. 10, no. 1, pp. 7–14, 2013.

[2] L. Hobert, A. Festag, I. Llatser, L. Altomare, F. Visintainer, and A. Kovacs, "Enhancements of v2x communication in support of cooperative autonomous driving," *IEEE Commun. Mag.*, vol. 53, no. 12, pp. 64–70, Dec. 2015.

[3] "Intelligent Transport Systems (ITS): Vehicular Communications; Basic Set of Applications; Definition," ETSI Std., Tech. Rep. ETSI TR 102 638, Jun. 2009.

[4] "5G automotive vision," White Paper, 5G PPP, Oct. 2015.

[5] M. Dohler, R. W. Heath, A. Lozano, C. B. Papadias, and R. A. Valenzuela, "Is the PHY layer dead?" *IEEE Commun. Mag.*, vol. 49, no. 4, pp. 159–165, Apr. 2011.

[6] J. Andrews, S. Buzzi, W. Choi, S. Hanly, A. Lozano, A. Soong, and J. Zhang, "What will 5g be?" *IEEE J. Sel. Areas Commun.*, vol. 32, no. 6, pp. 1065–1082, Jun. 2014.

[7] R. Balakrishnan and I. Akyildiz, "Local anchor schemes for seamless and low-cost handover in coordinated small cells," *IEEE Trans. Mobile Comput.*, vol. 15, no. 5, pp. 1182–1196, May 2016.

[8] C. L. I, J. Huang, R. Duan, C. Cui, J. . Jiang, and L. Li, "Recent progress on c-ran centralization and cloudification," *IEEE Access*, vol. 2, pp. 1030–1039, 2014.

[9] H. Zhang, C. Jiang, and J. Cheng, "Cooperative interference mitigation and handover management for heterogeneous cloud small cell networks," *IEEE Wireless Commun.*, vol. 22, no. 3, pp. 92–99, Jun. 2015.

[10] S. Y. Lien, S. C. Hung, K. C. Chen, and Y. C. Liang, "Ultra-low-latency ubiquitous connections in heterogeneous cloud radio access networks," *IEEE Wireless Commun.*, vol. 22, no. 3, pp. 22–31, Jun. 2015.

[11] R. Irmer, H. Droste, P. Marsch, M. Grieger, G. Fettweis, S. Brueck, H.-P. Mayer, L. Thiele, and V. Jungnickel, "Coordinated multipoint: Concepts, performance, and field trial results," *IEEE Commun. Mag.*, vol. 49, no. 2, pp. 102–111, February 2011.

[12] J.-M. Moon and D.-H. Cho, "Efficient handoff algorithm for inbound mobility in hierarchical macro/femto cell networks," *IEEE Commun. Lett.*, vol. 13, no. 10, pp. 755–757, Oct. 2009.

[13] J. Andrews, S. Singh, Q. Ye, X. Lin, and H. Dhillon, "An overview of load balancing in hetnets: old myths and open problems," *IEEE Wireless Commun.*, vol. 21, no. 2, pp. 18–25, Apr. 2014.

[14] J. Choi, W.-H. Lee, Y.-H. Kim, J.-H. Lee, and S.-C. Kim, "Throughput estimation based distributed base station selection in heterogeneous networks," *IEEE Trans. Wireless Commun.*, vol. 14, no. 11, pp. 6137–6149, Nov. 2015.

[15] X. Luo, "Delay-oriented QoS-aware user association and resource allocation in heterogeneous cellular networks," *IEEE Trans. Wireless Commun.*, vol. 16, no. 3, pp. 1809–1822, Mar. 2017.

[16] M. A. A. Masri and A. B. Sesay, "Mobility-aware performance evaluation of heterogeneous wireless networks with traffic offloading," *IEEE Trans. Veh. Technol.*, vol. PP, no. 99, pp. 1–1, 2015.

[17] F. Mehmeti and T. Spyropoulos, "Performance analysis of mobile data offloading in heterogeneous networks," *IEEE Trans. Mobile Comput.*, vol. 16, no. 2, pp. 482–497, Feb. 2017.

[18] T. M. Ho, N. H. Tran, L. B. Le, W. Saad, S. M. A. Kazmi, and C. S. Hong, "Coordinated resource partitioning and data offloading in wireless heterogeneous networks," *IEEE Commun. Lett.*, vol. 20, no. 5, pp. 974–977, May 2016.

[19] A. Apostolaras, G. Iosifidis, K. Chounos, T. Korakis, and L. Tassiulas, "A mechanism for mobile data offloading to wireless mesh networks," *IEEE Trans. Wireless Commun.*, vol. 15, no. 9, pp. 5984–5997, Sep. 2016.

[20] S. Paris, F. Martignon, I. Filippini, and L. Chen, "An efficient auction-based mechanism for mobile data offloading," *IEEE Trans. Mobile Comput.*, vol. 14, no. 8, pp. 1573–1586, Aug. 2015.

[21] H. Pilaram, M. Kiamari, and B. H. Khalaj, "Distributed synchronization and beamforming in uplink relay asynchronous ofdma comp networks," *IEEE Trans. Wireless Commun.*, vol. 14, no. 6, pp. 3471–3480, Jun. 2015.

[22] D. Lee, H. Seo, B. Clerckx, E. Hardouin, D. Mazzarese, S. Nagata, and K. Sayana, "Coordinated multipoint transmission and reception in LTE-advanced: deployment scenarios and operational challenges," *IEEE Commun. Mag.*, vol. 50, no. 2, pp. 148–155, Feb. 2012.

[23] R. Annavajjala, "Low-complexity distributed algorithms for uplink comp in heterogeneous lte networks," *IEEE J. Commun. Netw.*, vol. 18, no. 2, pp. 150–161, Apr. 2016.

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

[25] S. Bassoy, M. Jaber, M. A. Imran, and P. Xiao, "Load aware self-organising user-centric dynamic CoMP clustering for 5G networks," *IEEE Access*, vol. 4, pp. 2895–2906, 2016.

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

[27] F. Guidolin, L. Badia, and M. Zorzi, "A distributed clustering algorithm for coordinated multipoint in lte networks," *IEEE Wireless Commun. Lett.*, vol. 3, no. 5, pp. 517–520, Oct 2014.

[28] China Mobile Research Institute, "White paper of next generation fronthaul interface," Jun. 2015.

[29] W. Y. Lee and I. F. Akyildiz, "Spectrum-aware mobility management in cognitive radio cellular networks," *IEEE Trans. Mobile Comput.*, vol. 11, no. 4, pp. 529–542, Apr. 2012.

[30] M. Haenggi, J. Andrews, F. Baccelli, O. Dousse, and M. Franceschetti, "Stochastic geometry and random graphs for the analysis and design of wireless networks," *IEEE J. Sel. Areas Commun.*, vol. 27, no. 7, pp. 1029–1046, Sep. 2009.

[31] J. F. C. Kingman, *Poisson Processes*. Oxford University Press, 1993.

[32] D. P. Bertsekas and R. G. Gallager, *Data Networks*. Prentice Hall, 1992.

[33] L. Georgiadis, M. J. Neely, and L. Tassiulas, "Resource allocation and cross-layer control in wireless networks," *Foundations and Trends in Networking*, vol. 1, no. 1, pp. 1–144, 2006.

[34] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction*. The MIT Press, 1998.

[35] C. W. Sung and W. S. Wong, "User speed estimation and dynamic channel allocation in hierarchical cellular system," in *IEEE Vehicular Technology Conference (VTC)*, Jun. 1994, pp. 91–95 vol.1.

[36] M. Neely, "Energy optimal control for time-varying wireless networks," *IEEE Trans. Inf. Theory*, vol. 52, no. 7, pp. 2915–2934, Jul. 2006.

[37] C.-S. Chang and J. A. Thomas, "Effective bandwidth in high-speed digital networks," *IEEE J. Sel. Areas Commun.*, vol. 13, no. 6, pp. 1091–1100, Aug 1995.

[38] D. Wu and R. Negi, "Effective capacity: a wireless link model for support of quality of service," *IEEE Trans. Wireless Commun.*, vol. 2, no. 4, pp. 630–643, Jul. 2003.

[39] J. Choi, "Effective capacity of noma and a suboptimal power control policy with delay qos," *IEEE Trans. Commun.*, vol. 65, no. 4, pp. 1849–1858, Apr. 2017.

[40] A. A. Khalek, C. Caramanis, and R. W. Heath, "Delay-constrained video transmission: Quality-driven resource allocation and scheduling," *IEEE J. Sel. Topics Signal Process.*, vol. 9, no. 1, pp. 60–75, Feb. 2015.

[41] Y. Mao, J. Zhang, S. H. Song, and K. B. Letaief, "Stochastic joint radio and computational resource management for multi-user mobile-edge computing systems," *IEEE Trans. Wireless Commun.*, vol. PP, no. 99, pp. 1–1, 2017.

[42] 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.

[43] S.-C. Hung, S.-Y. Lien, and K.-C. Chen, "Stochastic topology cognition in heterogeneous networks," in *IEEE Personal, Indoor and Mobile Radio Communications (PIMRC)*, Sep. 2013, pp. 194–199.