IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

SPECIAL SECTION ON INTERNET OF THINGS (IOT) IN 5G WIRELESS COMMUNICATIONS

Received August 16, 2016, accepted October 3, 2016, date of publication November 3, 2016, date of current version March 28, 2017.

Digital Object Identifier 10.1109/ACCESS.2016.2624755

INVITED PAPER

A Survey of Client-Controlled HetNets for 5G

MICHAEL WANG[1,2], JIASI CHEN[3], EHSAN ARYAFAR[4], AND MUNG CHIANG[1]

[1]Department of Electrical Engineering, Princeton University, Princeton, NJ 08544, USA
[2]BAE Systems, Inc., Burlington, MA 01803, USA
[3]Department of Computer Science and Engineering, University of California at Riverside, Riverside, CA 92521, USA
[4]Intel Labs Santa Clara, CA 95054, USA

Corresponding author: M. Wang (michael.wang1@baesystems.com)

ABSTRACT With the rise and widespread deployment of a vast array of wireless radio access technologies (e.g., 3G, 4G/LTE, 802.11, Bluetooth, and Femto) and the growth of interest in potential 5G technologies such as millimeter-wave radio, coupled with the rapid increase in the number of network edge devices with multiple radio interfaces, the question of network control and client-to-base-station association becomes an important issue. Older, well-studied centralized control schemes where a single computational entity harvests channel information from individual clients in order to determine optimal resource allocations for each client is no longer tenable: such methods require significant signaling overhead which does not scale well with the expected number of hundreds of thousands of smart client devices with multiple radio interfaces capable of leveraging many different radio access technologies (RATs). With the rise of these smart devices, which come with significant computational power, it is now possible to ask the question: can the network allow the client control over RAT selection and association in order to meet some client-driven or network-driven objective, and to what degree does the network assist the client in making these choices? This question becomes particularly important given the increasing interest in standardization and deployment of client-controlled edge networking, or Fog networking. In this paper, we explore the spectrum of client-controlled HetNets for 5G networks: from the fully devolved distributed local control approach, where clients make local decisions without any assistance from the network, to the hybrid control approach where clients may make decisions given some global information provided by the network.

INDEX TERMS HetNets, 5G, cellular networks, distributed control, radio access networks.

I. INTRODUCTION

Heterogeneity of modern wireless network radio access technologies (RATs) (such as 3G, 4G/LTE, Wi-Fi, Bluetooth, and potential 5G technologies) is a critical component for ensuring network access and communication in current- and next-generation wireless networks. With so many different networking options available, and with modern mobile and edge devices sufficiently equipped with multiple wireless interfaces to take advantage of these different networks, these mobile devices are able to switch between different networks in an opportunistic way to perform services useful to the user.

However, this increased access to different networks (Fig. 1) comes with the added requirement and complexity of determining which network a client should connect with at any given time. In a chaotic radio environment (e.g. a bustling urban downtown neighborhood), channel conditions change so frequently that fine control of the clients in the network is required in order to prevent certain behaviors such as frequent switching between networks from damaging overall network performance. The main question to be answered in the Heterogeneous Networks (HetNets) scenario is **"How should a user select a RAT at any given time?"** For example, there are many efforts under way to specify solutions for networking cellular technologies (LTE, 3G, etc.) and IEEE 802.11 (Wi-Fi) technologies in the Third Generation Partnership Project (3GPP) [1], [2].

One traditional approach taken by network operators and academic research on HetNets is to allocate authority to a centralized agent or controller, which is then able to distribute edge devices and users over different networks based on some network objective such as load balancing. This approach has the advantage of finding a global optimal operating point

2169-3536 © 2016 IEEE. Translations and content mining are permitted for academic research only.
Personal use is also permitted, but republication/redistribution requires IEEE permission.
See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

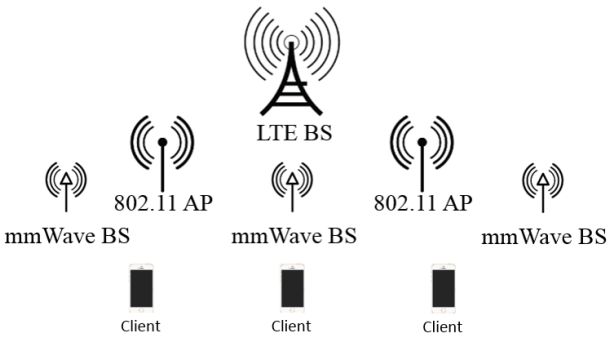

FIGURE 1. An example of a Heterogeneous Network with clients able to access 5G millimeter wave RATs

by assigning devices to RATs. However, with increasing numbers of mobile devices at the network edge, the message passing required from edge devices to the centralized agent to communicate available RATs and channel qualities for each device can quickly grow exponentially, resulting in an untenable situation for high-density areas like urban environments. Furthermore, different network operators seldom have an incentive to cooperate: for example, Boingo Wireless has no financial benefit in offloading its own WiFi customers onto Verizon's cellular network, and vice versa.

Meanwhile, with the advent of smartphones and other edge devices with high computational capacity, it is increasingly common to find devices released by device vendors with some form of association control given to the user, such as control loops that compare received signal powers from different RATs to select the optimal RAT. Localizing multiple aspects of the network association control plane to the edge device allows for a more accurate local view of not only the ambient radio environment around the client, but also the current applications that are running the client device, any user-specific preferences for data and energy usage, as well as the remaining battery life of the device itself. Furthermore, pressing issues present in a centralized decision-making scheme such as ownership and control of the intelligence for RAT selection over multiple BSs are avoided by allowing individual clients to determine which network they should connect to.

The heterogeneity (e.g. latency, bandwidth, packet loss, availability, etc.) of networks, coupled with the rise of increased computational power on modern mobile and edge devices, has increasingly led to the question of not only which RAT to select, but also where the intelligence for RAT selection in HetNets should be located, and it is now not inconceivable to place the some functionality for RAT association at the network edge–on the client device itself. In fact, with interest in client-driven network control (used in applications such as edge networking, or Fog networking [3]) increasing in the past few years, network-controlled centralized solutions can no longer be applied to problems such as client-controlled data transfer, storage and processing on the network edge.

In the paper we survey a variety of different client-centric approaches in localizing RAT selection and association for HetNets, and how they may be extended to be used with next-generation wireless technologies. Termed 5G or 5th generation wireless systems, these technologies are expected to provide data rates in the tens of Mbps, with speeds upwards of 1 Gpbs in special cases, as well as increased spectral efficiency, decreased latency, improved coverage and enhanced signalling [4]. Although HetNets is not an intrinsic part of the 5G definition [5], integrating 5G technologies with other existing deployments into a larger system of HetNets can provide a large payoff for both the network and the clients in terms of increased throughput, lower latency, better network load balancing, and other benefits.

The rest of the paper is structured as follows. We describe past approaches to HetNets and the client-centric HetNets model in Section II, and a baseline algorithm to solve the problem. Next, we discuss extensions to the basic model, recent work for each category as well as future work in Section III. We then discuss recent progress in industry and standards bodies for HetNets in 5G in Section IV. Finally, we discuss some fundamental assumptions made in HetNets in the conclusion in Section V.

II. NETWORK MODEL FOR 5G HetNets

A. HetNet SELECTION

There have been many different approaches to modeling the RAT selection problem in HetNets [6]–[14], where clients are allowed to switch radio interfaces for data transmission in order to improve some metric (Fig. 2a). These approaches have attempted to ask and answer the questions of *where* in the wireless network that RAT selection should take place, and *how* RAT selection should take place: specifically, what algorithm should be used to determine which RATs to connect to.

Traditionally, solutions to these questions have been primarily network-centric, setting *where* to be inside the network core, and *how* to be some centralized agent running an optimization algorithm that makes a global client-to-Base-Station/eNode-B association decision for all clients in the network at the same time. The advantages of such an architecture, where all decisions are made on the network side, are that the RAT selection algorithm can act upon a big-picture view of the network (can poll clients for their local channel conditions and have loading information on the RATs themselves) and converge to and achieve a globally-optimal result with respect to some network metric relatively quickly. This approach is generally favored by organizations such as cellular service providers that seek to maintain control of network operation, reflecting the view that the network should optimize for some network-wide metric that the service provider wishes to focus on.

In network-controlled HetNets, these decisions are made in the following way (Fig. 2b): first, the centralized agent that performs the decision-making polls all clients seeking to connect to one or more of the RATs controlled by the agent,

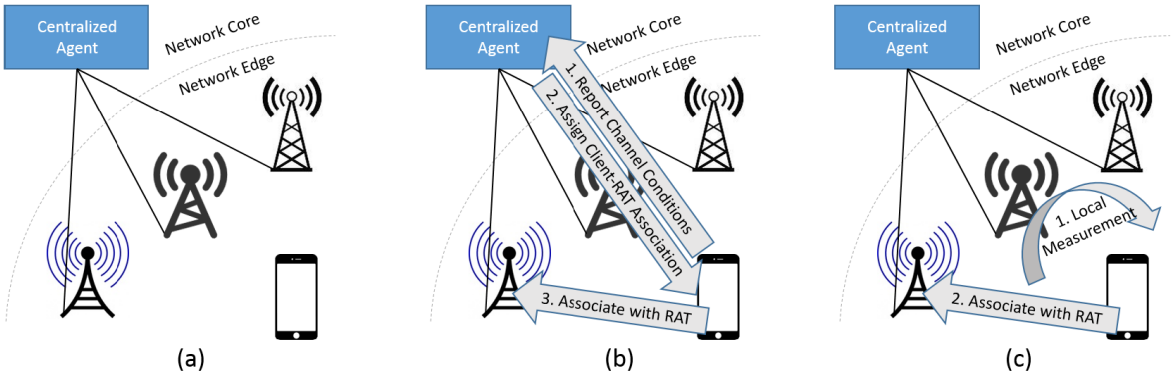

FIGURE 2. (a) Client with access to 3 Heterogeneous RATs; (b) Network-controlled HetNets: (1) The client reports information on its accessible RATs to the centralized agent, which (2) computes the globally optimal client-RAT association back to the client and (3) switches RATs; (c) Client-controlled HetNets: (1) The client gathers information on its accessible RATs and makes a local decision for its own client-RAT association decision and (2) switches RATs.

requesting information on the client's local channel conditions. Based on this information, it calculates the optimal association of clients to RATs, and then transmits these associations back to each client. Once the client receives these instructions (by way of the RAT it is currently using), it then changes its RAT association as directed.

The exact optimization performed is unique to each network. Depending on the type of network and the network operator, different networks may be designed to optimize for different metrics depending on their business needs:

- Maximize aggregate throughput and fairness of resource allocation [8], [9], [15]
- Load balancing [10]–[12], [16]
- Minimize outage probability [13], [14]
- Maximize a measure of Quality of Service

A variety of techniques are used to achieve these goals such as cost functions, utility maximization, stochastic geometry and combinatorial optimization to determine the "best" [17] network to associate with [18] and [19].

However, there are downsides to centralized network-controlled HetNets. Chief among these are the issues of timeliness of switching, scalability of the system to multiple clients with multiple interfaces, and the issue of how to obtain global control of client-RAT association. Any centralized scheme would have to obtain client inputs (e.g., channel conditions, battery life, number and type of interfaces on the device, application type, etc.) in order to optimally allocate resources. Such polling would incur significant delay in each calculation due to latency between client and controller. Furthermore, all clients need to convey this information to the network–depending on the geographical area that the centralized agent controls, the sheer amount of required overhead traffic simply for passing clients' parameters to the network could become a significant portion of data that the network transports, further limiting the efficiency of the HetNet. Finally, different RATs belonging to different business entities may not be willing to pool resources and allow outside control of proprietary networks. For example, to achieve true centralized network-controlled HetNets, companies such as Verizon and Boingo must be willing to hand off their existing customer traffic to each other when instructed to do so–raising questions about fairness of traffic allocation between their respective cellular and Wi-Fi networks, how to agree upon a method of traffic allocation, and even customer privacy.

In contrast to centralized network-controlled HetNets, client-controlled HetNets focuses solely on the clients' perspective. Based on local observations at the client itself, the client must make a decision to associate with the RAT that provides some optimal metric (e.g., a utility function, throughput, etc.), evaluated locally (Fig. 2c). Although this form of distributed optimization run on individual clients may not obtain a client-RAT association with as good a metric as the centralized network-controlled case [20], [21], there are several advantages. First, the timeliness problem of the centralized case is avoided: all measurements do not suffer from network latency, as all measurements are locally available on the client. Second, scalability of RAT selection in HetNets is maintained as each client only needs to calculate its optimal association. Third, client privacy is preserved as it no longer needs to transfer information about its traffic (data type, utility function, battery power, etc.) to agents in the network. Finally, the different business entities needing to share control of their networks is avoided–as long as each client has permission to associate with a RAT, each network only need allow the client to connect when requested.

B. NETWORK MODEL

Next, we describe the basic model for client-controlled HetNets, where the user controls its own RAT selection decisions based on its local view of the network.

Let M be the set of Base Stations (BS) accessible to the set of distinct client devices in a given physical area, N, some of which may be 5G wireless technologies, where M the number of BS available and N the number of

TABLE 1. Main notation.

$\mathbf{M}$	Set of RATs	$\mathbf{N}$	Set of clients
M	Number of RATs	N	Number of clients
U_i	Utility function for client i	$\omega_{i,k}$	Throughput of client i on RATs k
$X_{i,k}$	Binary-valued Association Variable	n_k	Number of clients on RATs k
K_i	Single-/Multi-homed parameter	η	Switching threshold
T	minimum measurement intervals before switching	p	Randomization Parameter
h	Hysteresis parameter for switching	m_i	Backoff memory
$R_{i,k}$	PHY Rate of client i on BS k		

clients in the area. Here, Base Station (BS) is used as a generic term to represent millimeter-wave BS in 5G mmWave, eNB in 4G, AP in 802.11, etc., and all BSs are assumed to be non-interfering due to frequency reuse or spatial separation between same-RAT BSs and frequency separation between different-RAT BSs. Each client device can simultaneously communicate with a subset of BSs up to the number of RAT interfaces it is equipped with.[1]

The goal of the network is to determine how to best choose which RAT a client should connect to at any given time. The general RAT selection problem in HetNets has the following form: maximize some aggregate utility function over all individual clients (Eq. 1), subject to some association constraints on the clients (Eq. 2,3) over the decision variables (Eq. 5):
maximize

$$\sum_i U_i\left(\omega_{i,1}, \omega_{i,2}, ..., \omega_{i,M}\right) \ \forall i \in N \tag{1}$$

subject to

$$\sum_k X_{i,k} \leq K_i \forall i \in N, k \in M \tag{2}$$

$$\omega_{i,k} = f(Y_{1,k}, \ldots, Y_{N,k}) \forall i \in N, k \in M \tag{3}$$

$$Y_{i,k} = X_{i,k} \cdot R_{i,k} \forall i \in N, k \in M \tag{4}$$

with variables

$$X_{i,k} \in \{0, 1\} \quad \forall i \in N, k \in M \tag{5}$$

where $U_i(.)$ is the utility function being maximized, $K_i = 1$ for single-homed clients ($K_i > 1$ if multi-homing is allowed), $\omega_{i,k}$ is the throughput of client i on RAT k, which depends on a specific throughput sharing function $f(Y_{1,k}, ..., Y_{N,k})$, and $Y_{i,k} = X_{i,k} \cdot R_{i,k}$ is the product of the PHY-layer rate $R_{i,k}$ of client i on RAT k and the binary-valued association variable $X_{i,k}$ ($X_{i,k} = 1$ if client i associates with RAT k and 0 otherwise).

These utility functions may be designed to optimize different metrics ranging from weighted aggregate throughput and cost of obtaining service, to fairness of obtained throughput and network load balancing.

In client-centric RAT selection in HetNets, this problem is solved by placing the RAT association decision on each of the individual clients with varying degrees of assistance provided by the BS. Each client makes a locally-optimal decision based on knowledge of its radio environment, and act in a distributed manner. In such client-centric approaches, showing the system is stable (i.e., it converges) is key–otherwise an algorithm may lead to compounding negative behaviors such as infinite oscillation and a suboptimal outcome for all clients involved.

[1]Due to frequency separation, it is assumed that each RAT interface can observe signals from at most one BS at any time, and for interfaces that receive signals from multiple BSs, the functionality is assumed to be modeled by treating each BS as a distinct RAT: e.g. an 802.11 interface with Channels 1,6,11 accessible viewed as a 3-RAT interface.

C. BASELINE CLIENT-CENTRIC RAT SELECTION

The problem of RAT selection is modeled as a noncooperative game [20] in which clients select BSs to maximize each client's individual throughput in a distributed manner. In this game formulation, the player set is defined as the set of users N, and their strategies are the set of accessible BSs M.

In [20], all RATs are classified into one of two classes of throughput-sharing models that are functions of the individual PHY-layer rates ($R_{i,k}$) for each client associating with the RAT ($X_{i,k} > 0$). The first class of models is characterized by throughput sharing functions that depend on the maximum rates of clients on the RAT $R_{i,k}$ (e.g. proportional-fair), with Eq. 6 replacing Eq. 3:

$$\omega_{i,k} = f_k(Y_{1,k}, Y_{2,k}, ..., Y_{N,k}) \quad \forall i \in N \tag{6}$$

where $f_k(.)$ is a RAT-specific throughput-sharing function that depends on $Y_{i,k} = X_{i,k} \cdot R_{i,k}$, the PHY-layer rates of all clients associated with RAT k at that time. An example of this is the downlink coordination function (DCF) in 802.11 that provides fair access:

$$\omega_{i,k} = \frac{L}{\sum_j \frac{L}{Y_{j,k}}} \quad \forall i \in N \tag{7}$$

where L is the packet length (e.g. downlink WiFi throughput). With the assumption that each client can obtain or predict $R_{i,k}$ simple knowledge of the type of throughput-sharing model can result in accurate prediction of potential obtainable throughput ($\omega_{i,k}(t)$ at each time t).

The second class of models is characterized by throughput sharing functions $f_k(.)$ that depend only on the total number of users that share the RAT (n_k), and not all of the clients' individual rates. This model allows for individual client throughputs to be distinct from that of other clients sharing the same RAT (e.g. time-fair), with Eq. 8 replacing Eq. 3:

$$\omega_{i,k} = Y_{i,k} \times f_k(n_k) \quad \forall i \in N \tag{8}$$

where $n_k = |Y_{i,k} > 0|$ is the number of clients on RAT k. An example of this type of sharing function is time-fair TMDA MAC protocols, where each client has an equal amount of time to transmit:

$$\omega_{i,k} = \frac{Y_{i,k}}{n_k} \ \forall i \in N \tag{9}$$

Algorithm 1 Baseline Client-Centric RAT Selection Algorithm

Input: user i's parameters: η, T, p, h, Set of RATs
Output: Decision to switch, and the selected RAT

```
1  for each RAT k' do
2      if ω_{i,k'}[t+1] / ω_{i,k}[t] > η, ∀ t = t − T + 1, ..., t then
3          if class(k') = class(k) then
4              if rand < p^{m_i+1} then
5                  switch to k'
6                  if concurrent move then increment m_i
7
8                  else reset m_i to 0
9
10         else
11             if ω_{i,k'} > h then
12                 if rand < p^{m_i+1} then
13                     switch to k', update h
14                     if concurrent move then
                       increment m_i
15
16                     else reset m_i to 0
17
```

D. BASELINE CLIENT-CENTRIC ALGORITHM

The baseline algorithm for client-centric RAT selection in [20] is shown below in Algorithm 1, where the separable objective in Eq. 1 is maximized individually by each client seeking to maximize its own throughput $\omega_{i,k}$. In order for a client i to switch at time $t+1$ from BS k to BS k', the expected throughput gain defined as $\frac{\omega_{i,k'}[t+1]}{\omega_{i,k}[t]}$ should exceed a threshold (η) for the past T time slots (Line 2), where T corresponds to the frequency of measurement. If multiple clients simultaneously switch to the same BS, they would observe a mismatch between their observed throughputs and their predicted values. To minimize concurrent switches to the same BS, clients are assumed to switch probabilistically with probability $p < 1$ (Line 4). The randomization parameter, p, depends on the congestion in the network and acts similarly to the 802.11 contention window mechanism. However, when simultaneous switching to a BS happens, all clients involved set their randomization parameter to p^{m_i+1} (Line 6), as in binary exponential backoff, where m_i is the number of past consecutive concurrent switch observed by i.

As clients locally determine which RAT to associate with in order to maximize their own local utilities, it may be that some clients continually switch without convergence. Hysteresis is introduced to dampen oscillations so that the system converges to equilibrium. This parameter, h, represents the dependence of RAT selection to the historical behavior of the client and its previous switches. Algorithm 1 shows an example policy where a client is only allowed to change between BSs of different classes if it has an expected throughput greater than its hysteresis value (Lines 8-9).

This algorithm is guaranteed to converge to a Nash Equilibrium for any combination of BSs from either class (Theorems $1-3$, [20]). Furthermore, if all BSs are time-fair, the average Pareto-efficiency gain (average per-client throughput improvement) between a non-Pareto-optimal Nash output of Alg 1 and a Pareto-dominant profile is bounded by 2 if $N \leq M$ and $\frac{N+M}{N}$ if $M \geq N$ (Theorem 6, [20])). Similarly, if all BSs are proportional-fair, then the same average Pareto-efficiency gap is bounded by $2 \cdot (1 + \log(N))$ if $N \leq M$ and $\frac{N+M}{N} \cdot (1 + \log(N))$ otherwise (Theorem 6, [20]).

III. EXTENSIONS OF CLIENT-CENTRIC RAT SELECTION

A. HYBRID CONTROL

Hybrid control extends the concept of client-centric control of RAT selection in HetNets to allow for network-assistance in the RAT selection process. This type of shared control preserves the client's right to make the final decision on when and where to switch, but gives the network some input in the switching itself, such as providing information on network metrics such as load balancing. This added information is useful because the purely-local view of the network observed by fully-distributed client-centric RAT selection is not guaranteed to be accurate. Operating under this limited perspective can lead to suboptimal behavior when a client simply ignores the effect of its presence upon other users. Several works address this by allowing the network to inform clients with some global knowledge such as [21]–[24] for example.

A broadcast technique to inform clients making local decisions of network conditions is used in several approaches [22], [23]. In [22], the authors designed a BS association system for HetNets in which base stations broadcast both their current weighted load and price. These parameters are then used by clients to select and associate to the base station that best satisfies their utility. Under this service model, clients are assisted to make the best decision even under mobility, and can even assign individual applications to different interfaces.

This is taken one step further in [23] where the authors develop a low-complexity distributed algorithm that uses gradient descent dual-decomposition to split the multi-homed joint RAT association problem into two distinct subproblems of identifying the optimal BS at the client, and updating a BS-specific multiplier to be broadcast after each iteration at the BS. The Lagrangian multiplier acts as the price of the

BS determined by its load, and performs a sort of load-balancing in the network itself.

The authors in [24] develop a distributed algorithm in which the clients do not directly compete to maximize throughput–instead, they associate to BSs in a way to maximize the total reward obtained from the BS in order to prevent selfish behavior. The reward assigned by each BS is dependent on the loss of throughput incurred for other users due to association based on *marginal cost pricing* [25].

Theoretical results for generalized distributed client-centric RAT selection in HetNets with prioritized service were analyzed in [21]. Both a purely client-centric and the hybrid association model were analyzed, and showed that purely client-centric association with generic weights can result in infinite oscillations; but under several specific classes of weighted priorities, convergence can be guaranteed for the system. Tight polynomial and linear bounds are found for the client-centric model, and that the proper selection of a potential function by the network can guarantee convergence of the system.

Assignment problem approaches for traffic offloading in HetNets for femto have also been proposed [26], [27] that find optimal association. Compared to the above works, matching schemes explicitly allow for indirect negotiation between clients and BSs in allocating network resources. Instead of providing a price parameter to clients, these approaches directly rank clients for each BS in terms of how well a potential BS may maximize the client or network metric.

In [26], the problem of achieving proportional-fair throughput for client-RAT association in HetNets is transformed into an equivalent matching problem, which can be solved in polynomial time. This exchange, where the BSs iteratively announce their price of association, and clients submit bids for resources, results in a global reduction for macrocellular traffic of up to 30% compared to several non-cooperative game-based strategies.

A preference list approach is used in [27] to find a stable matching between clients and femtocells for *uplink*. In this setup, both clients and BSs rank each other based on preference functions that capture clients' utilities which depend on packet success rate, delay, and small cells' incentive to extend macrocell coverage. The game is solved using two phases involving admission games that allow transfers between BSs, followed by data transmission, with performance improvements of up to 23% compared to a best packet success rate algorithm.

Although the network may provide some additional information, the ultimate decision to switch still rests on the client itself–and in the absence of perfect, global knowledge, the client will need to discover how it should associate with the BSs to meet its objective. Online learning techniques such as reinforcement learning [28] and multi-armed bandits [29] are among some of the methods used to explore and exploit the spectrum resources that are accessible to a client.

Q-learning is used in [28] to learn the client-specific bias values for received power in order to perform cell-range expansion for HetNets. By using these individualized bias values over a common bias over all UEs, the system is a multi-agent system that leverages distributed learning where information is never shared. The costs are reported back to the clients from the BSs and are used to update the Q-values for future ranking of RATs for association.

B. FAST TIME VARYING RATs (mmWAVE)

Several white papers by several industry groups predit [4], [5], 5G wireless systems will provide, at the very least, higher data rates (between 10 Gbps in an indoor office environment to 25 Mbps in a very dense crowd, with an average of 50 Mbps for general use), end-to-end latency on the order of 1 ms, and handle up to 150, 000/km^2 connections in a crowd-like environment.

The key elements of 5G identified by many telecom companies are [30]–[32]:

- **Peak Data Rate**: 10 Gbps per client, 4x that of 4G
- **End-to-End Latency**: 1 ms, 1/50 that of 4G
- **Scale of Connections**: 1 million/km^2, 100x that of 4G

In order to meet these goals, 5G will not only require access to diverse (licensed, shared licensed, and unlicensed) spectrum [32], [33], but leverage new spectrum bands from 400 MHz to 100 GHz [34], [35] and to create capability to handle dense HetNets [36].

Millimeter-wave (mmWave) radio technologies is increasingly looked at as one of the primary new RATs for 5G given the scarcity of spectrum at microwave frequencies [37]–[39]. A combination of cost-effective hardware, high-gain and steerable antennas, larger carrier bandwidth allocations all translate to higher data rates for mmWave communications for small-cell and indoor applications on the order of 200m [40], [41]. However, mmWave is also characterized by higher path-loss exponents and an inability to penetrate obstructions such as the human body.

There exist many studies on mmWave wireless communication in the 60 GHz band [42]–[46]. These studies characterize the free space propagation loss and the higher loss due to attenuation from non-Line-of-Sight (NLOS) channels. However, these bands may not be suitable for cellular due to the unlicensed nature of this band and the coexistence of 802.11ad, and the primary candidates are 28, 38, 71-76 and 81-86 GHz for indoor applications [47].

Outage studies for 38 GHz were done in Texas in 2012 [48] and the first statistical channel models were measured for 28 and 73 GHz in dense urban neighborhoods (New York City) in 2013 [49], which analyzed a range of parameters: path loss, shadowing, mmWave line-of-sight(LOS)/NLOS/Outage probabilities, angular orientation of the BSs and clients, and number of clusters. Furthermore, the authors showed strong evidence for the existence of a third state, Outage, characterized by a complete loss of signal, and proposed modeling mmWave with a three-state channel model defined by the degrees of signal obstruction.

mmWave technology is particularly sensitive to obstruction, with drastic changes in the path-loss exponent for

different mmWave frequency bands [39]. These increase from values of 1.8 and 2 for 28 GHz and 73 GHz with a direct line-of-sight component to values upwards of 4.5 and 2.69 once the direct path is blocked. Furthermore, human-body blockage can have a significant effect on signal quality, with attenuation between 20-35dB [50], [51] with the loss of the line-of-sight.

However, with exceedingly high throughput upwards of 10+ Gbps for 73 GHz demonstrated only in the past year [52], a large path-loss exponent only helps the case that these high throughput mmWave BSs can be used as dense small cells in HetNets to handle extremely high throughput and high volume traffic.

In [29], the authors apply online learning using multi-armed bandits to RAT Selection for HetNets with fast-changing mmWave channels to maximize throughput while minimizing parameterized switching costs, with optimal total regret. Limited feedback from the BS in the form of a parameter describing the channel state between client and BS is sent to each client, and the client then uses this knowledge of the channel to discover and exploit the ''best" BS in terms of average throughput using a upper-confidence-bound-type algorithm in which the client is only allowed to switch BS associations at specific times. An alternative approach that also leverages the Markovian nature of channel state changes is considered in [53], which directly considers dynamic channel load and link quality but not switching costs. However, the use of mmWave in HetNets is a relatively new area of study, and there remain lots of work to be done to better understand how to interchangeably use mmWave alongside other existing networks.

C. NOISY INFERENCE OF CLIENT METRICS

In order for client-centric BS association algorithms to function correctly, they require a method for differentiating one BS from another to determine which is optimal for association. These metrics depend on the concerns of the client, which may vary for different applications (e.g. maximizing throughput for bandwidth-hungry video or minimizing end-to-end latency for web applications). These metrics are highly sensitive to noisy estimates, and can become a bottleneck to optimal association of clients to BSs. In the ideal situation, this information could be accurately inferred at the client given some additional knowledge, or provided by the BS that performs some asynchronous calculation; however, in the general case this cannot be assumed.

Inaccurate inference can result in a variety of inefficiencies [20]. First, oscillations may result if a client can frequently switch between two different valuations of distinct BSs, causing it to repeatedly associate to one BS only to switch to the other. This frequent change in available throughput and load on the BSs in question can have an adverse impact on other clients in the network, resulting in a cascading effect where one oscillating client can cause oscillations to propagate throughput the network as other clients see their throughputs and latencies change due to incorrect switching. Next, incorrect inference may not result in optimal achieved metrics for the client: inference errors may result in inefficient or incorrect BS associations, leading to suboptimal metrics. Furthermore, additional costs can be incurred: switching between different BSs (and indeed different types of RATs as well) require that the client and network set up a new connection and tear down the old one in a handover process which consumes both time that could be otherwise used for downloading/uploading data, as well additional battery. For such power-limited devices such as smartphones and mobile devices, it is highly undesirable to allow noisy estimates and inferences to cause oscillations and incorrect switches.

One solution to this is to adapt the decision threshold for switching (e.g. η in [20]) in a client-centric control algorithm to the ambient noise in the inference. By learning the distribution of inference noise on each BS, it is possible to negate its impact (e.g. by subtracting the worst-case empirically observed error) on the ranking of the BS. By increasing the required gain in predicted metrics required to initiate a switch between BSs, the client can switch more conservatively. However, the system becomes less likely to switch with an increased decision threshold–and it is less likely to make smaller corrections in client-BS association that would increase the overall efficiency of the network. Decision thresholds that control switching in client-centric control algorithms can be used as a control knob–a tuning mechanism that controls both convergence speed and resiliency to noisy metric inference. By increasing the switching threshold, clients switch less frequently and are more resilient to noise–but lose out on optimality.

D. MULTIHOMING

In addition to single RAT selection, a plurality of RATs may be selected and multiplexed by the client. Compared to single RAT selection, the multi-homed scenario is much more challenging because in addition to deciding *which* RATs should be used, an additional decision of *how much* each RAT should be used must be made (assuming finite backlog of traffic). In the context of Section II-B, the network association variable $X_{i,k}$ is no longer integer, and constraint in Eq. 5 instead becomes:

$$0 \leq X_{i,k} \leq 1 \tag{10}$$

and the number of RATs each user can connect to $K_i > 1$. While at first glance, the multi-homing problem may seems simpler since the variables are no longer integer, solving the problem can still be challenging depending on the form of the throughput model (e.g., Eq. 7). Moreover, additional practical issues such as throughput estimation and feedback, protocol design, and application-specific performance must also be considered.

As a specific example of the multi-homed problem formulation, consider the case of video streaming, which dominates data traffic on today's Internet. Video streaming differs file transfer or web browsing in that video is encoded and streamed at a (roughly) constant rate. The large video file is usually split into several smaller pieces, or ''chunks'',

TABLE 2. Comparison of approaches to multi-homing.

Name	Goal	Layer	Inputs	Control Knobs	Analysis framework
LIA (MPTCP) [54]	congestion control	transport	RTT, packet loss	congestion window	heuristic
OLIA [55]	congestion control	transport	RTT, packet loss, bits received	congestion window	fluid model, Pareto optimality
Balia [56]	congestion control	transport	RTT, packet loss	congestion window	fluid model, utility maximization
SCTP [57]	congestion control	transport	RTT, packet loss	congestion window	heuristic
miDRR [58]	enable user interface preferences	application	throughput, user preferences	subflow selection	generalization of deficit round-robin
MicroCast [59]	cooperative video downloading	application, network	user requests	download schedule, network coding, pseudo-broadcast	network coding
ATOM [60]	balance WiFi and LTE usage	control plane	throughput, data plan costs	user association	utility maximization

which are downloaded at regular intervals from the server. Whenever the client requests a video chunk, the video chunk should be downloaded quickly in order to avoid video stalls and meet the playback deadline. The question, then, is what fraction of the chunk should be requested on each RAT, in order to satisfy the video rate requirement C (in bits) while minimizing the chunk download time? The model in Section II-B is flexible and can accommodate this by defining the utility function as:

$$U_i = -\max_k \left\{ \frac{X_{i,k} C}{\omega_{i,k}} \right\} \tag{11}$$

with new constraint:

$$\sum_k X_{i,k} = 1 \tag{12}$$

(11) defines the utility as a function of the video chunk download time across all RATs, i.e., we are minimizing download time so that the video buffer grows and the chance of stalling decreases. (12) says the sum of the fractional allocations of C is equal to one. The variable $X_{i,k}$ is the fraction of bits sent over network k.

Next, we survey some recent works on multi-homing, which we also summarize in Table 2. In the transport layer, a major standardization effort towards enabling multi-homing on the Internet today is multipath-TCP (MPTCP) [61]. MPTCP creates multiple subflows out of a single TCP flow, each of which can be bound to a different RAT. This pooling of resources can enable higher throughputs, easier mobile handovers, and improved path diversity for failure recovery; but the challenges include backwards-compatibility with regular TCP and overcoming middle-boxes who do not recognize or permit MPTCP options. The IETF has standardized MPTCP [61], but measurement studies thus far have shown limited adoption in the public Internet [62].

To perform rate adaptation for client-controlled HetNets, several MPTCP congestion control algorithms [54]–[56] have been proposed. Specifically, the main control mechanism is the per-subflow congestion window, which changes the sending rate of each subflow based on the congestion of each RAT. Various objectives have been proposed, such as TCP-friendliness [54], Pareto-optimality [55], and utility maximization [56]. The sending rate is further complicated by scheduling algorithms that run on top of congestion control, breaking ties if there are multiple subflows with space in the congestion window [63]. In both congestion control and scheduling, each client uses local measurements of congestion (e.g., RTT), so the approach is distributed, but some work has been done on proving convergence to a globally optimal solution [56].

Other transport-layer approaches apart from MPTCP have also been considered, such as [64], which uses multiple single-path TCP connections on multiple RATs and develops Markov models for analysis. [65] proposes an approach that is particularly geared towards wireless networks with packet losses, and uses explicit congestion notification (ECN) as well as forward error correction (FEC) to recover from losses.

While transport-layer protocols tend to optimize for traditional QoS metrics such as throughput, latency, and loss, application layer-based multi-homing can consider additional factors such as economic cost and content sharing. Application-layer solutions, while not standardized, may enable more rapid adoption, since end-to-end transport-layer protocols need not be modified. Several creative uses for multi-homed devices have been proposed. Reference [58] considers scheduling for mobile devices where the user can explicitly specify RAT preferences (e.g., user does not want to use LTE due to monetary cost), and applications can explicitly specify which RATs may be used (e.g., video requires

RATs with high throughputs). The resulting scheduler is a generalization of deficit round-robin. Reference [66] studies rate allocation in the multi-user case, based on video characteristics and network measurements of throughput and latency. Their optimal solution depends on knowing video packet distortion characteristics, while their practical solution is based on H^{∞} control theory. Reference [60] studies multi-homing from the perspective of a cellular operator who manages both cellular and WiFi networks. The operator wishes to assign users to WiFi and LTE networks to both maximize utility and minimize cost, and takes a centralized optimization approach. Reference [59] proposes using each RAT for different functionality: the cellular connection is used to download videos, and WiFi is used to share the videos amongst a group. A combination of network coding and pseudo-broadcasting help ensure that the content is delivered reliably and efficiently.

On the implementation side, [67] demonstrates an Android prototype of multi-homing using virtual interfaces, similar to the MPTCP design. Reference [68] focuses on video streaming and prototypes a custom YouTube player that integrates with the existing YouTube library. MPTCP has also released open-source Linux kernels for desktops, Android phones, routers, and cloud services.

E. Wi-Fi OFFLOADING

Opportunistic client-centric switching based on local availability of non-cellular networks in HetNets has already been well-studied in the context of 802.11 Wi-Fi over different timescales (sub-second to tens of seconds and more) [69]–[71], and much of it remains applicable to HetNets in a 5G environment. Wi-Fi networks have been shown to be a high capacity option for offloading traffic from cellular networks, accepting 65% of total mobile client traffic while saving 55% of the battery by only offloading data to Wi-Fi when the network is available [69].

The Wiffler system [70] has also shown up to a 45% reduction in cellular workload for data with a delay tolerance of up to 60 seconds. By leveraging delay tolerance of different types of data and fast switching, Wiffler is able to route data from one RAT to another if the first BS is unable to satisfy the traffic delay requirements. The system is able to opportunistically leverage the offloading capacity of Wi-Fi networks (if present), and fall back upon other traditional cellular networks if no Wi-Fi networks that can satisfy data delay requirements is within reach.

The authors in [71] have gone further, developing client-stored models of daily mobility in order to perform predictive forecasts for the mobile client's radio environment. By leveraging the habitual behavior of people taking similar paths during their daily lives and combining past wireless measurements, a system for determining typical Wi-Fi BS quality and client location can generate connectivity forecasts of which Wi-Fi BSs can be available to in the future, and can assist in opportunistic client-centric data offloading, though the system requires some training.

Many works also address Wi-Fi offloading in the presence of some network prediction. This has been studied for integration of cellular-and-WLAN networks [72], which supports network discovery and selection to help clients discover non-cellular networks. In [73], HotZones, uses prediction to download delay-tolerant content when close to Wi-Fi BSs. It creates a rank-ordered list of most frequently visited BS based on past client behavior. This profile is shared with the network operator, and the total aggregate list is broadcast to all clients–in effect, bypassing the high overhead of opportunistic scanning for higher-throughput Wi-Fi RATs. Clients may then connect to various Wi-Fi BSs as they wish with the added knowledge of a measure of load on those BSs, and are found to be able to offload up to 70% of their cellular traffic to those BSs.

Large-scale city traces of mobile clients over the course of 30 days were used in MADNet [74] to evaluate the gains of citywide (San Francisco) Wi-Fi offloading using metropolitan BSs. It allows the cellular network to reduce load by signaling over cellular, but performing download/uploading over Wi-Fi based on explicitly defined client preferences. More than half of cellular traffic was offloaded, and file transfer delay was reduced by more than 50% in the majority of requests.

Many of these client-centric approaches remain valid for HetNets with 5G, as the ubiquitous deployment of Wi-Fi for both indoor and outdoor coverage and offloading is not expected to be replaced anytime soon. These techniques may also be extended for unmanaged 5G RATs deployed as small cells in both licensed and unlicensed spectrum, however these techniques must be updated so that they function on the faster timescale of 5G mmWave RATs (e.g. order of milliseconds) in an efficient way. Opportunistic use of these RATs in the home or office environment within a single room or a floor has the potential to maintain the benefits of load balancing and coverage of Wi-Fi offloading, but also to exploit the higher throughput potential of 5G technologies such as mmWave.

F. GAME THEORETIC ANALYSIS

Noncooperative game theory, in which individual clients make local decisions to maximize individual payoffs without a means to enforce restrictions on the behavior of other clients, is often used for fully-distributed coordination without any sort of management from a non-client party–which is the case for Distributed RAT Selection. In this model, the set of players is the set of clients, and the set of player strategies is the set of BSs (or RATs) that they may associate with at a given time. A game of this type is said to have *converged* to a Nash Equilibrium (NE) if each player considers its selected strategy to be optimal given the choices of all other players–that is, it cannot unilaterally improve its payoff by changing strategies.

A common class of techniques for distributed coordination is found in the area of noncooperative congestion games [75], [76], where players select from a common set

of strategies to play and the reward of each strategy is a monotonically nonincreasing function of the total number of players playing that strategy [77]. In [75], by considering an entire type of RAT to be a single BS in a congestion game framework, a finite improvement path can be found for the congestion game where each asynchronous client can selfishly switch to reduce their cost until a NE is reached. However, to implement this, the authors caution that to reach a pure NE requires exact information on incurred cost, which may be difficult to obtain in a timely and accurate manner. The authors in [76] model downlink access to multiple broadband BSs as a congestion game, which models the client- and BS-specific cost of association as the congestion impact on other clients sharing the same network. By abstracting away the multi-rate property of HetNets, tight analytical bounds for the price of anarchy and price of stability (ratio between the value of the ''best/worst" equilibria points and an optimal solution) are found.

Evolutionary games, in which groups of clients select strategies to play against clients from other groups, have also been considered [78], [79]. In [78], a population game for multihomed RAT association is studied for 802.11 under evolutionary dynamics. Prices based on channel occupancy and total throughput in the cell are used to calculate payoff functions for each client to calculate a potential function for the population game, and it is shown that the stationary points of such a game are asymptotically stable and maximizes throughput. In [79], an evolutionary game is used to perform client-driven network load balancing between different types of networks (specifically WMAN, cellular, and WLAN). Two solutions to obtaining evolutionary equilibrium are presented. The population evolution solution relies on coordination between clients to share knowledge of the average payoff in a given area, so that underperforming clients may change networks; and reinforcement learning leverages Q-learning to explore and rank the different networks for optimal empirical payoff.

G. PROBABILISTIC ANALYSIS

Markov Decision Processes (MDP) have also been used to model the HetNets RAT selection problem at the client side [80]–[83]. Clients may internally store empirical knowledge on the rewards (e.g. throughputs) obtained in each state (e.g. BS) that are accessible, as well as their transition probabilities. Every time period, the client must make a decision on which action to take–which RAT or BS to associate with, in order to maximize some expected total reward.

[80] develop a MDP model for vertical handoff between different types of RATs, that considers the link reward for connecting to a BS, the signaling load and processing load when the handoff is performed. The algorithm relies solely on implicit feedback from the network in the link reward obtained after connecting to the BS, and shows improvement over several other algorithms for vertical handoff.

Markov chains can also be used to explicitly model ongoing voice and data sessions on individual RATs [81]. With the assumption that calls and data sessions begin and end sequentially (they arrive/depart individually) and that the total traffic offered to both networks are known, this work develops a 4D Markov Chain to model two TDMA and WCMDA networks in order to simulate the performance of several RAT selection policies.

Application-specific models can also be used, such as for Video-on-Demand [82]. In this work, multihomed clients optimize their choice of RATs for the minimization of video playback disruption costs and the communication cost of receiving a video chunk over a given RAT. The MDP is used to determine which RAT to send a chunk request to at a given time, and the resulting adaptive ATAC policy is shown to have lower costs than policies with static thresholds for request allocation.

IV. CURRENT STATE OF THE INDUSTRY

The goal of HetNets is to enable the seamless association, data-transfer, BS switching, and dissociation for a client to a set of wireless networks of different types, in a way that maximizes the aggregate utility of the clients and networks involved. With such a vast number of different types of RATs, each with different specifications, effective transmit/receive ranges, supported data-rates, and speed of channel changes, finding a simple unifying algorithm for HetNets has been a question that many organizations in industry have considered.

There are several obstacles to developing a common approach to the integration of new technologies such as 5G into the HetNets architecture, such as noise, fast temporal variations, hybrid control schemes, multi-homing and load balancing. Recently, industry groups such as 3GPP and IEEE have been pressing forward with standardization initiatives to enable HetNets for existing technologies, as well as laying the groundwork for next-gen 5G technologies such as mmWave.

In Release 12 [84], multiple enhancements were released to improve client mobility in HetNets for both LTE and UMTS, including cell discovery (client-based discovey, network-based discovery and collaborative client and network-based discovery), and general enhancements for small cell deployments. Furthermore, 3GPP has included standardization for dual connectivity (simultaneous multihoming to both macro- and microcells by clients) [85], allowing for dynamic traffic routing over multiple paths.

However, on the core philosophy of where in the network HetNets control should reside, the debate is still ongoing in 3GPP, between centralized solutions that place client-BS association decisions in the network, distributed solutions that place client-BS association decisions strictly on the client, and the hybrid approach, where the association decision is made by the network guiding clients in the association process. Several different techniques have been proposed [86], ranging from client-centric BS selection with network assistance (broadcasts) and with network assistance and network-advised policies, as well as network-controlled BS selection with network-determined policies.

Without a solid determination of *who* should control BS selection and *where* that intelligence shall be placed, many industry leaders in the consumer electronics world have been moving forward with their own homegrown approaches to managing access to HetNets. In the past, Apple Inc. had built their flagship smartphone, the iPhone, such that it automatically offloaded the client onto Wi-Fi wherever possible in order to decrease use of the client's subscribed data plan. Furthermore, in early 2016, Apple introduced Wi-Fi Assist, a feature that detects a poor Wi-Fi signal and automatically switches the mobile device back to using data from your cellular plan [87].

When the Wi-Fi Assist feature was released, however, there was a large backlash against Apple because many device owners were either uninformed or unaware of the feature which resulted in individuals incurring thousands of dollars worth of data overage charges due to inadvertent data use [88]. This approach, although done in the best interest of maximizing quality of service by finding the BS with the best signal, clearly demonstrates the overall failure of the system when it is designed to optimize an objective or metric that doesn't match that of the client (in this case, total monetary cost was not considered) even though the BS association ostensibly gave full control to the user.

V. CONCLUSION

As the growth of computational power on the mobile client device grows, it is increasingly possible to leverage the techniques discussed in this survey work to place the intelligence for RAT selection on the network edge, literally in the hands of the user. With massive amounts of machine-type communication and the rise of IoT on the horizon, traditional centralized decision schemes where a centralized controller gathers channel information from all users about their channel conditions and local radio environment is no longer tenable for the massive scale that can be expected from the rise of these new networking trends for smart home, smart office, and smart everything.

Therefore, new network selection approaches must be developed for HetNets in the context of 5G that can take advantage of the capabilities available on the network edge–some of that work has already begun. However, many questions still remain to be studied in detail:

- **Network Assistance for Client-Centric HetNets.** To what degree should the client be autonomous in selecting a RAT for association? There is a vast gulf between fully-distributed approaches and hybrid solutions that leverage a degree of network assistance.
- **Objective Formulation.** How should the objective be designed for client-centric RAT association? Many approaches involve network-centric or client-centric objectives, but perhaps a combination of the two can balance the needs of both the network and the client.
- **Performance Gap.** How worse off will these client-centric solutions be compared with a centralized implementation with global knowledge? Can this gap be quantified in terms of Price of Anarchy or Stability?
- **Timescales.** The temporal variability of the channel conditions of newer RATs (e.g. mmWave) may be different from existing technologies: How should RAT Selection algorithms account for this difference in timescales of channel variation?

Furthermore, there have been additional criteria that tend to be assumed in discussions on HetNets–but are not necessarily guaranteed. These criteria simplify analysis, but cannot be assumed in the general case:

- **Quota.** The mobile device's data quota is assumed to be infinite, and it always has a positive marginal utility for downloading or uploading more data.
- **Battery.** The battery capacity of the mobile device is assumed to be infinite, and many of the works in this survey do not consider the joint problem of RAT selection under explicit finite power constraints.
- **WiFi Coverage.** Smartphones and mobile devices are assumed to always have access to an alternative network, and it is always possible to access both cellular and Wi-Fi networks, which may not be true in certain scenarios (e.g. rural settings).

ACKNOWLEDGMENTS

The authors would like to thank Nageen Himayat, Sarabjot Singh, Shu-Ping Yeh, Shilpa Talwar and David Ott at Intel Labs for helpful discussions, as well as the Intel "Higher, Denser, Wilder" 5G program in conjunction with the University of Southern California and New York University.

REFERENCES

[1] *3GPP/WLAN RAN Interworking, Release 12*, Standard 3GPP TR 37.834, Jan. 2014.
[2] *WLAN/3GPP Radio Interworking*, Standard 3GPP RP-122038, Dec. 2012.
[3] M. Chiang and T. Zhang. (Jan. 2016). *Fog Networking: An Overview on Research Opportunities*. [online] Available: http://arxiv.org/abs/1601.00835.
[4] *5G White Paper*, NGMNAlliance, 2015, pp. 1–125. [Online]. Available: https://www.ngmn.org/uploads/media/NGMN_5G_White_Paper_V1_0.pdf
[5] *Understanding 5G: Perspectives on Future Technological Advancements in Mobile*, GSMAIntelligence, London, U.K., 2014, pp. 1–26.
[6] M. Kassar, B. Kervella, and G. Pujolle, "An overview of vertical handover decision strategies in heterogeneous wireless networks," *Comput. Commun.*, vol. 31, no. 10, pp. 2607–2620, Jun. 2008.
[7] X. Yan, Y. A. Şekercioğlu, and S. Narayanan, "A survey of vertical handover decision algorithms in fourth generation heterogeneous wireless networks," *Comput. Netw.*, vol. 54, no. 11, pp. 1848–1863, Aug. 2010.
[8] Y. Lin, W. Bao, W. Yu, and B. Liang, "Optimizing user association and spectrum allocation in HetNets: A utility perspective," *IEEE J. Sel. Areas Commun.*, vol. 33, pp. 1025–1039, Jun. 2015.
[9] D. Bethanabhotla, O. Y. Bursalioglu, H. C. Papadopoulos, and G. Caire, "Optimal user-cell association for massive MIMO wireless networks," *IEEE Trans. Wireless Commun.*, vol. 15, no. 3, pp. 1835–1850, Mar. 2016.
[10] S. Singh, H. S. Dhillon, and J. G. Andrews, "Downlink rate distribution in multi-RAT heterogeneous networks," in *Proc. IEEE Int. Conf. Commun. (ICC)*, Jun. 2013, pp. 5188–5193.
[11] J. G. Andrews, S. Singh, Q. Ye, X. Lin, and H. S. Dhillon, "An overview of load balancing in HetNets: Old myths and open problems," *IEEE Wireless Commun.*, vol. 21, no. 2, pp. 18–25, Apr. 2013.
[12] D. Bethanabhotla, O. Y. Bursalioglu, H. C. Papadopoulos, and G. Caire, "User association and load balancing for cellular massive MIMO," in *Proc. Inf. Theory Appl. Workshop (ITA)*, Feb. 2014, pp. 1–10.

[13] H. S. Jo, Y. J. Sang, P. Xia, and J. G. Andrews, "Outage probability for heterogeneous cellular networks with biased cell association," in *Proc. IEEE Global Telecommun. Conf. (GLOBECOM)*, Dec. 2011, pp. 1–5.

[14] H. ElSawy and E. Hossain, "Two-tier HetNets with cognitive femtocells: Downlink performance modeling and analysis in a multichannel environment," *IEEE Trans. Mobile Comput.*, vol. 13, no. 3, pp. 649–663, Mar. 2014.

[15] S. Corroy, L. Falconetti, and R. Mathar, "Dynamic cell association for downlink sum rate maximization in multi-cell heterogeneous networks," in *Proc. IEEE Int. Conf. Commun. (ICC)*, Jun. 2012, pp. 2457–2461.

[16] A. Chakraborty, V. Navda, V. N. Padmanabhan, and R. Ramjee, "Coordinating cellular background transfers using loadsense," in *Proc. 19th Annu. Int. Conf. Mobile Comput. Netw. (MobiCom)*, 2013, pp. 63–74.

[17] E. Gustafsson and A. Jonsson, "Always best connected," *IEEE Wireless Commun.*, vol. 10, no. 1, pp. 49–55, Feb. 2003.

[18] L. Wang and G. S. Kuo, "Mathematical modeling for network selection in heterogeneous wireless networks-Ů-A tutorial," *IEEE Commun. Surveys Tuts.*, vol. 15, no. 1, pp. 271–292, 1st Quart., 2013.

[19] D. Liu *et al.*, "User association in 5G networks: A survey and an outlook," *IEEE Commun. Surveys Tuts.*, vol. 18, no. 2, pp. 1018–1044, 2nd Quart., 2015.

[20] E. Aryafar, A. Keshavarz-Haddad, M. Wang, and M. Chiang, "RAT selection games in HetNets," in *Proc. IEEE INFOCOM*, Apr. 2013, pp. 998–1006.

[21] E. Monsef, A. Keshavarz-Haddad, E. Aryafar, J. Saniie, and M. Chiang, "Convergence properties of general network selection games," in *Proc. IEEE Conf. Comput. Commun. (INFOCOM)*, Apr. 2015, pp. 1445–1453.

[22] S. Deb, K. Nagaraj, and V. Srinivasan, "MOTA: Engineering an operator agnostic mobile service," in *Proc. 17th Annu. Int. Conf. Mobile Comput. Netw. (MobiCom)*, 2011, pp. 133–144.

[23] Q. Ye, B. Rong, Y. Chen, M. Al-Shalash, C. Caramanis, and J. G. Andrews, "User association for load balancing in heterogeneous cellular networks," *IEEE Trans. Wireless Commun.*, vol. 12, no. 6, pp. 2706–2716, Jun. 2012.

[24] P. Coucheney, C. Touati, and B. Gaujal, "Fair and efficient user-network association algorithm for multi-technology wireless networks," in *Proc. IEEE INFOCOM*, Apr. 2009, pp. 2811–2815.

[25] R. Cole, Y. Dodis, and T. Roughgarden, "Pricing network edges for heterogeneous selfish users," in *Proc. 35th Annu. ACM Symp. Theory Comput. (STOC)*, 2003, pp. 521–530.

[26] W. Wang, X. Wu, L. Xie, and S. Lu, "Femto-matching: Efficient traffic offloading in heterogeneous cellular networks," in *Proc. IEEE Conf. Comput. Commun. (INFOCOM)*, Apr. 2015, pp. 325–333.

[27] W. Saad, Z. Han, R. Zheng, M. Debbah, and H. V. Poor, "A college admissions game for uplink user association in wireless small cell networks," in *Proc. IEEE Conf. Comput. Commun. (INFOCOM)*, Apr. 2014, pp. 1096–1104.

[28] T. Kudo and T. Ohtsuki, "Cell range expansion using distributed *Q*-learning in heterogeneous networks," *J. Wireless Commun. Netw.*, vol. 2013, no. 1, pp. 1–10, Mar. 2013.

[29] M. Wang, A. Dutta, S. Buccapatnam, and M. Chiang, "Regret-minimizing exploration in HetNets with mmWave," in *Proc. IEEE SECON*, Jun. 2016, pp. 1–9.

[30] Ericsson. (May 2016). *5G for The Networked Society*. [Online]. Available: http://www.ericsson.com/spotlight/5g

[31] Huawei. (Apr. 2016). *The Road to 5G*. [Online]. Available: http://www.huawei.com/minisite/5g/en/

[32] Qualcomm, "Building a unified 5G platform: For the next decade and beyond," Qualcomm, San Diego, CA, USA, Tech. Rep., Rep. no, 2015.

[33] InterDigital. (Apr. 2016). *The Road to 5G*. [Online]. Available: http://www.huawei.com/minisite/5g/en/

[34] Nokia Networks, "5G master plan," Espoo, Finland, Tech. Rep., Rep. no, 2015.

[35] Nokia Networks, "5G radio access: system design aspects," Espoo, Finland, Tech. Rep., Rep. no, 2015.

[36] 4G Americas, "5G spectrum recommendations," Bellevue, WA, USA, Tech. Rep., Rep. no, 2015.

[37] F. Boccardi, R. W. Heath, A. Lozano, T. L. Marzetta, and P. Popovski, "Five disruptive technology directions for 5G," *IEEE Commun. Mag.*, vol. 52, no. 2, pp. 74–80, Feb. 2014.

[38] T. E. Bogale and L. B. Le. (Oct. 2015). "Massive MIMO and millimeter wave for 5G wireless HetNet: Potentials and challenges." [Online]. Available: https://arxiv.org/abs/1510.06359

[39] Y. Niu, Y. Li, D. Jin, L. Su, and A. V. Vasilakos, "A survey of millimeter wave communications (mmWave) for 5G: Opportunities and challenges," *Wireless Netw.*, vol. 21, no. 8, pp. 2657–2676, 2015.

[40] T. S. Rappaport *et al.*, "Millimeter wave mobile communications for 5G cellular: It will work!" *IEEE Access*, vol. 1, pp. 335–349, May 2013.

[41] (Jul. 2015). *Nokia Networks, Nyu Wireless Host Brooklyn 5G Summit To Advance A Super-Fast Generation Of Mobile Communications*. [Online]. Available: http://nyuwireless.com/newevents/nokia-networks-nyu-wireless-host-brooklyn-5g-summit-to-advance-a-super-fast-generation-of-mobile-communications/

[42] T. S. Rappaport, J. N. Murdock, and F. Gutierrez, Jr., "State of the art in 60-GHz integrated circuits and systems for wireless communications," *Proc. IEEE*, vol. 99, no. 8, pp. 1390–1436, Aug. 2011.

[43] S. Geng, J. Kivinen, X. Zhao, and P. Vainikainen, "Millimeter-wave propagation channel characterization for short-range wireless communications," *IEEE Trans. Veh. Technol.*, vol. 58, no. 1, pp. 3–13, Jan. 2009.

[44] F. Giannetti, M. Luise, and R. Reggiannini, "Mobile and personal communications in the 60 GHz band: A survey," *Wireless Pers. Commun.*, vol. 10, no. 2, pp. 207–243, Jul. 1999.

[45] H. Xu, V. Kukshya, and T. S. Rappaport, "Spatial and temporal characteristics of 60-GHz indoor channels," *IEEE J. Sel. Areas Commun.*, vol. 20, no. 3, pp. 620–630, Apr. 2002.

[46] R. C. Daniels, J. N. Murdock, T. S. Rappaport, and R. W. Heath, Jr., "60 GHz wireless: Up close and personal," *IEEE Microw. Mag.*, vol. 11, no. 7, pp. 44–50, Dec. 2010.

[47] A. Ghosh *et al.*, "Millimeter-wave enhanced local area systems: A high-data-rate approach for future wireless networks," *IEEE J. Sel. Areas Commun.*, vol. 32, no. 6, pp. 1152–1163, Jun. 2014.

[48] J. N. Murdock, E. Ben-Dor, Y. Qiao, J. I. Tamir, and T. S. Rappaport, "A 38 GHz cellular outage study for an urban outdoor campus environment," in *Proc. IEEE Wireless Commun. Netw. Conf. (WCNC)*, Apr. 2012, pp. 3085–3090.

[49] M. R. Akdeniz *et al.*, "Millimeter wave channel modeling and cellular capacity evaluation," *IEEE J. Sel. Areas Commun.*, vol. 32, no. 6, pp. 1164–1179, Jun. 2014.

[50] J. S. Lu, D. Steinbach, P. Cabrol, and P. Pietraski "Modeling human blockers in millimeter wave radio links," *ZTE Commun.*, vol. 10, pp. 23–28, Dec. 2012.

[51] M. Gapeyenko *et al.*, "Analysis of human-body blockage in urban millimeter-wave cellular communications," in *Proc. IEEE Int. Conf. Commun. (ICC)*, 2016, pp. 1–7.

[52] M. Alleven. (2015). "Nokia networks paves way for 5G with 10 GBPS demo at brooklyn 5G summit." [Online]. Available: http://engineering.nyu.edu/in-media/2015/04/10/nokia-networks-paves-way% -5g-10-gbps-demo-brooklyn-5g-summit

[53] M. Mezzavilla, A. Dhananjay, S. Panwar, S. Rangan, and M. Zorzi. (2015). "An MDP model for optimal handover decisions in mmWave cellular networks." [Online]. Available: https://arxiv.org/abs/1507.00387

[54] C. Raiciu *et al.*, "How hard can it be? designing and implementing a deployable multipath TCP," in *Proc. 9th USENIX Conf. Netw. Syst. Design Implementat.*, 2012, pp. 399–412.

[55] R. Khalili, N. Gast, M. Popovic, and J. Y. L. Boudec, "MPTCP is not Pareto-optimal: Performance issues and a possible solution," *IEEE/ACM Trans. Netw.*, vol. 21, no. 5, pp. 1651–1665, Oct. 2013.

[56] Q. Peng, A. Walid, J. Hwang, and S. H. Low, "Multipath TCP: Analysis, design, and implementation," *IEEE/ACM Trans. Netw.*, vol. 24, no. 1, pp. 596–609, Feb. 2016.

[57] *Stream Control Transmission Protocol*, document RFC 4960, IETF, 2016.

[58] K.-K. Yap, T.-Y. Huang, Y. Yiakoumis, S. Chinchali, N. McKeown, and S. Katti, "Scheduling packets over multiple interfaces while respecting user preferences," in *Proc. ACM CoNEXT*, 2013, pp. 109–120.

[59] L. Keller, A. Le, B. Cici, H. Seferoglu, C. Fragouli, and A. Markopoulou, "Microcast: Cooperative video streaming on smartphones," in *Proc. ACM MobiSys*, 2012, pp. 57–70.

[60] R. Mahindra, H. Viswanathan, K. Sundaresan, M. Y. Arslan, and S. Rangarajan, "A practical traffic management system for integrated LTE-WiFi networks," in *Proc. ACM MobiCom*, 2014, pp. 189–200.

[61] A. Ford, C. Raiciu, M. Handley, S. Barre, and J. Iyengar. (2011). *Architectural Guidelines for Multipath TCP Development*. [Online]. Available: https://www.rfc-editor.org/info/rfc6182

[62] O. Mehani, R. Holz, S. Ferlin, and R. Boreli, "An early look at multipath TCP deployment in the wild," *Proc. HotPlanet*, pp. 7–12, Sep. 2015.

[63] C. Paasch, S. Ferlin, O. Alay, and O. Bonaventure, "Experimental evaluation of multipath TCP schedulers," in *Proc. ACM SIGCOMM Workshop Capacity Sharing Workshop*, 2014, pp. 27–32.
[64] B. Wang, W. Wei, Z. Guo, and D. Towsley, "Multipath live streaming via TCP: Scheme, performance and benefits," *ACM Trans. Multimedia Comput. Commun. Appl.*, vol. 5, pp. 1–23, Aug. 2009.
[65] V. Sharma, K. Kar, K. K. Ramakrishnan, and S. Kalyanaraman, "A transport protocol to exploit multipath diversity in wireless networks," *IEEE/ACM Trans. Netw.*, vol. 20, no. 4, pp. 1024–1039, Aug. 2012.
[66] X. Zhu, P. Agrawal, J. Pal Singh, T. Alpcan, and B. Girod, "Rate allocation for multi-user video streaming over heterogenous access networks," in *Proc. ACM Multimedia*, 2007, pp. 37–46.
[67] K.-K. Yap *et al.*, "Making use of all the networks around us: A case study in Android," in *Proc. ACM CellNet Workshop*, 2012, pp. 19–24.
[68] Y.-C. Chen, D. Towsley, and R. Khalili, "Msplayer: Multi-source and multi-path leveraged youtuber," in *Proc. ACM CoNEXT*, 2014, pp. 263–270.
[69] K. Lee, J. Lee, Y. Yi, I. Rhee, and S. Chong, "Mobile data offloading: How much can WiFi deliver?" *IEEE/ACM Trans. Netw.*, vol. 21, no. 2, pp. 536–550, Apr. 2013.
[70] A. Balasubramanian, R. Mahajan, and A. Venkataramani, "Augmenting mobile 3G using WiFi," in *Proc. 8th Int. Conf. Mobile Syst., Appl., Services (MobiSys)*, 2010, pp. 209–222.
[71] A. J. Nicholson and B. D. Noble, "Breadcrumbs: Forecasting mobile connectivity," in *Proc. 14th ACM Int. Conf. Mobile Comput. Netw. (MobiCom)*, 2008, pp. 46–57.
[72] *Access Network Discovery and Selection Function*, document TS 24.312 v11.5.0, 3GPP, Dec. 2012.
[73] N. Ristanovic, J. Y. L. Boudec, A. Chaintreau, and V. Erramilli, "Energy efficient offloading of 3G networks," in *Proc. IEEE 8th Int. Conf. Mobile Ad-Hoc Sensor Syst.*, Oct. 2011, pp. 202–211.
[74] S. Dimatteo, P. Hui, B. Han, and V. O. K. Li, "Cellular traffic offloading through WiFi networks," in *Proc. IEEE 8th Int. Conf. Mobile Ad-Hoc Sensor Syst.*, Oct. 2011, pp. 192–201.
[75] M. Ibrahim, K. Khawam, and S. Tohme, "Congestion games for distributed radio access selection in broadband networks," in *Proc. IEEE Global Telecommun. Conf. (GLOBECOM)*, Dec. 2010, pp. 1–5.
[76] M. Cesana, N. Gatti, and I. Malanchini, "Game theoretic analysis of wireless access network selection: Models, inefficiency bounds, and algorithms," in *Proc. 3rd Int. Conf. Perform. Eval. Methodologies Tools*, 2008, p. 6.
[77] I. Milchtaich, "Congestion games with player-specific payoff functions," *Games Economic Behavior*, vol. 13, no. 1, pp. 111–124, 1996.
[78] S. Shakkottai, E. Altman, and A. Kumar, "Multihoming of users to access points in WLANs: A population game perspective," *IEEE J. Sel. Areas Commun.*, vol. 25, no. 6, pp. 1207–1215, Aug. 2007.
[79] D. Niyato and E. Hossain, "Dynamics of network selection in heterogeneous wireless networks: An evolutionary game approach," *IEEE Trans. Veh. Technol.*, vol. 58, no. 4, pp. 2008–2017, May 2009.
[80] E. Stevens-Navarro, Y. Lin, and V. W. S. Wong, "An MDP-based vertical handoff decision algorithm for heterogeneous wireless networks," *IEEE Trans. Veh. Technol.*, vol. 57, no. 2, pp. 1243–1254, Mar. 2008.
[81] X. Gelabert, J. Pérez-Romero, O. Sallent, and R. Agustí, "A Markovian approach to radio access technology selection in heterogeneous multiaccess/multiservice wireless networks," *IEEE Trans. Mobile Comput.*, vol. 7, no. 10, pp. 1257–1270, Oct. 2008.
[82] J. Lee and S. Bahk, "On the MDP-based cost minimization for video-on-demand services in a heterogeneous wireless network with multihomed terminals," *IEEE Trans. Mobile Comput.*, vol. 12, no. 9, pp. 1737–1749, Sep. 2013.
[83] X. Wang, J. Chen, A. Dutta, and M. Chiang, "Adaptive video streaming over whitespace: SVC for 3-tiered spectrum sharing," in *Proc. IEEE Conf. Comput. Commun. (INFOCOM)*, Apr. 2015, pp. 28–36.
[84] *Overview of 3GPP Release 12 Version 0.2.0*, document 3GPP TR 37.834, 2015.
[85] K. Mallinson. (2014). *Industry Voices*. [Online]. Available: http://www.3gpp.org/news-events/3gpp-news/1619-hetnet2
[86] *Study in Wireless Local Area Network (WLAN)—3GPP Radio Interworking Release 12*, document 3GPP 37.834, 2014.
[87] Apple Inc. (2016). *About Wi-Fi Assist*. [Online]. Available: https://support.apple.com/en-us/HT205296
[88] CBS News. (2016). *Apple's Wi-Fi Assist Feature Blamed for Teen's $2,000 Phone Bill*. [Online]. Available: http://www.cbsnews.com/news/apple-iphone-wi-fi-assist-blamed-for-teens-2000-phone-bill/

MICHAEL WANG received the dual B.S. degrees in physics and electrical engineering from The Pennsylvania State University (University Park) in 2011, and the Ph.D. degree in electrical engineering from Princeton University in 2016, under the supervision of Prof. M. Chiang of the EDGE Lab. He is currently a Principal Engineer with BAE Systems. His research interests range from RAT selection in HetNets to distributed control and deadline-aware resource allocation in the cloud.

JIASI CHEN received the Ph.D. degree from Princeton University and the B.S. degree from Columbia University. She is currently an Assistant Professor with the Department of Computer Science and Engineering, University of California at Riverside, Riverside. Her research interests include mobile wireless networks, video streaming, network economics, and the Internet-of-things, through a combination of mathematical analysis and systems development.

EHSAN ARYAFAR received the B.S. degree in electrical engineering from the Sharif University of Technology, Iran, in 2005, and the M.S. and Ph.D. degrees in electrical and computer engineering from Rice University, Houston, TX, USA, in 2007 and 2011, respectively. He was a Postdoctoral Research Associate with Princeton University from 2011 to 2013. He is currently a Research Scientist with Intel Labs. His research interests are in the areas of mobile and wireless networks, and span both algorithm design as well as system prototyping.

MUNG CHIANG is currently the Arthur LeGrand Doty Professor of Electrical Engineering with Princeton University. His textbook '*Networks: Friends, Money and Bytes* and online course reached 250 000 students since 2012. He founded the Princeton EDGE Lab in 2009, which bridges the theory-practice gap in edge networking research by spanning from proofs to prototypes. He co-founded a few startups in mobile, Internet of Things, and big data areas and co-founded the Open Fog Consortium. He is the Director of Keller Center for Innovations in engineering education at Princeton University and the inaugural Chairman of Princeton Entrepreneurship Council. His research on networking received the 2013 Alan T. Waterman Award, the highest honor to U.S. young scientists and engineers.