On WiFi Offloading in Heterogeneous Networks: Various Incentives and Trade-Off Strategies

Yejun He, *Senior Member, IEEE*, Man Chen, *Student Member, IEEE*,
Baohong Ge, and Mohsen Guizani, *Fellow, IEEE*

***Abstract*—Due to the rapid development of wireless access technologies and smart terminals, mobile data traffic is continuously increasing, which is expected to lead to an explosive growth of data in heterogeneous networks especially cellular networks. It is significant for network operators to expand the capacity of cellular networks to avoid congestion and overload so as to guarantee users' satisfaction. Given that contemporary terminals are capable of both WiFi and cellular networks, WiFi offloading is envisioned as a promising solution to utilize the various benefits of WiFi and cellular networks to migrate traffic from cellular networks to WiFi networks. This paper surveys the state-of-the-art progress in the field of WiFi offloading. After discussing the requirements from the emerging 5G technology regarding the coexistence of WiFi and cellular networks, selecting and switching schemes are presented. The bandwidth and capacity of WiFi networks are usually excellent, whereas the coverage and energy efficiency may be unacceptable. We elaborate on several existing solutions of WiFi offloading schemes and discuss how the parameters of several kinds of heterogeneous networks affect the offloading decision. We also illustrate how multiple networks cooperate in heterogeneous networks in order to balance the offloading performance. We classify current various incentives of WiFi offloading into five categories: 1) capacity; 2) cost; 3) energy; 4) rate; and 5) continuity. Improving the capacity is the basic incentive, which can be further classified in terms of delay techniques. From operators' and users' perspectives, we also investigate various state-of-the-art incentives of WiFi offloading such as minimizing cost, saving energy consumption, and improving rate. Furthermore, WiFi offloading schemes that attempt to enhance continuity to deal with frequent disruption problems are further investigated, especially for vehicular scenarios. Finally, future research directions and challenges for WiFi offloading strategies are presented in various incentives of WiFi offloading.**

***Index Terms*—WiFi, WiFi offloading, heterogeneous network, offloading incentive, network selection, vehicular data offloading.**

Manuscript received August 20, 2015; revised December 28, 2015 and March 2, 2016; accepted April 8, 2016. Date of publication April 25, 2016; date of current version November 18, 2016. This work was supported in part by the National Natural Science Foundation of China under Grant 61372077, in part by the Science and Technology Innovation Commission of Shenzhen under Grant JCYJ 20120817163755061 and Grant ZDSYS 201507031550105, and in part by the Guangdong Provincial Science and Technology Program under Grant 2013B090200011.

Y. He, M. Chen, and B. Ge are with the Shenzhen Key Laboratory of Antennas and Propagation, College of Information Engineering, Shenzhen University, Shenzhen 518060, China (e-mail: heyejun@126.com; starbank@126.com; gebaohong@163.com).

M. Guizani is with the Electrical and Computer Engineering Department, University of Idaho, Moscow, ID 83844, USA (e-mail: mguizani@ieee.org).

Digital Object Identifier 10.1109/COMST.2016.2558191

I. Introduction

MOBILE data traffic is rapidly increasing at an unprecedented rate with the proliferation of smart devises, which is known as the explosion of data traffic. Due to the development of wireless access technologies, cellular networks are capable of transmitting data traffic at high rates. Applications and traffic are gradually migrated from traditional Internet to wireless networks for the convenience of mobility without being concerned with transmitting rates. Researchers from Cisco announced that global mobile traffic grew 74% in 2016 [1]. Furthermore, Cisco predicts that the monthly global mobile data traffic will surpass 30.6 exabytes by 2020. In fact, as Cisco predicts, not only an increasing number of smart phones and tablets but also the emerging machine to machine (M2M) modules will contribute to the explosion of data traffic. The increase in mobile data traffic will be mostly generated by smart phones, while over two-thirds of the total traffic will be video and audio data.

To address the explosion of mobile data traffic, one solution is to upgrade existing networks to the next generation networks. Another solution is to increase the number of base stations and make the cell smaller to increase the capacity of cellular networks. The trouble is, these solutions require a huge value of capital expenditure (CAPEX) and operation expense (OPEX). There is still a hypothetical solution leveraging usage-based price plan, but it constrains data usage. This constraint does not conform to the future flat structure that is independent of usage. It is foreseeable that future networks are heterogeneous in nature [2].

Nevertheless, it is practical to use existing radio access technologies (RATs) to steer data traffic in heterogeneous networks from a comprehensive perspective. A scenario of coexistence in heterogeneous networks is shown in Fig. 1, where user equipment (UE) is capable of accessing various RATs. It is very interesting for the industry and academia to offload the data traffic from a cellular network to femtocells [3], WiFi, and more recently device-to-device opportunistic networks [4], which are referred to as mobile data offloading. Given that WiFi access points (APs) are currently widely deployed by operators and residents, WiFi offloading is envisioned as a promising solution to utilize the various benefits of WiFi and cellular networks. In fact, WiFi offloading has been increasingly deployed by a large number of mobile network operators (e.g., Verizon, AT&T, T-Mobile, and Vodafone) [5]. Thus, we limit the scope of this survey to WiFi offloading in particular.

TABLE I
DEFINITION OF INCENTIVES

Incentive	Definition from Operator	Definition from User
Capacity	The total WiFi offloaded mobile data traffic amount	Each WiFi offloaded UE throughput
Cost	A part of revenue except for CAPEX and OPEX	Data service fee paid by each user
Energy	The whole power consumption of cellular network	Each UE battery power consumption
Rate	The average completion time of demanded data transmissions	Data rate for each UE
Continuity	The frequency of network disconnections	The frequency of each UE disconnections

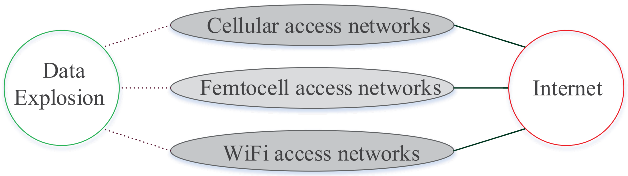


Fig. 1. To meet data explosion from UEs in heterogeneous networks, WiFi offloading migrates partial data traffic to WiFi access networks to alleviate congestion of cellular access networks. Offloading data to Femtocells is another way.

However, some problems need to be addressed for WiFi offloading. First, an appropriate scheduling scheme in the heterogeneous network is needed. It should be capable of managing its radio accessing and allow UE to select the access network with the best performance in terms of various practical requirements. The key is to utilize various factors to determine specific network for UE, achieving the best performance in terms of various incentives. This is related to some issues such as network selection scheme (NSS), and cooperative management in heterogeneous networks. Second, with different incentives of augmenting WiFi offloading schemes, the main challenges and methods may be quite different. Striving to improve one performance of an incentive may damage other performance of another incentive. For example, if there is an incentive to improve capacity to offload mobile data traffic to WiFi APs as far as possible [6], WiFi network congestion can occur and in turn can reduce cellular operators' revenue. Thus, the key challenge to augmenting WiFi offloading is to find out the essential factors that can be well utilized to achieve different tradeoffs between various performance metrics, according to different incentives.

For the convenience of our readers, we illustrate a typical trade-off. Delayed offloading scheme allocates some delay-tolerant data traffic to WiFi access networks, while allocating real-time data traffic to cellular networks. However, it is apparently not a wise choice to extend the delay to improve capacity without any limits. To ensure service quality, delay must be constrained by some mechanisms so that the delay does not exceed the limit of user's tolerance. The key is to make a tradeoff between delay and capacity. Some researchers attempt to obtain the tradeoff from experimental measurements, and others formulate the WiFi offload problem into an optimization problem. Similarly, some tradeoff problems may occur when we attempt to save cost for cellular operators, or save energy for UEs, etc. Thus, it is very wise to distinguish WiFi offloading according to different incentives for WiFi offloading to well investigate WiFi offloading schemes.

A. Motivation

To end chaotic state and avoid confusion, we attempt to present a guidance to help researchers quickly find their interested directions on WiFi offloading. We classify the state-of-the-art in this field into five categories, considering what they want to mainly improve and address. We call the metric of this classification as incentive as shown in Table I. These categories consist of capacity, cost, energy, rate, and continuity. That is because that the factors between which we need to trade off are quite different for different incentives. In fact, the key to augmenting WiFi offloading schemes for different incentives is usually to optimize the tradeoff strategies. For example, to save energy consumption for users, the factors between which we need to trade off include capacity, energy consumption, and quality of service, etc.

As introduced above, capacity is the most basic incentive of WiFi offloading to meet the explosion of data traffic. In detail, the incentive of improving capacity will be further classified into two subcategories: amount of the total mobile data traffic being offloaded, and the throughput per user (of LTE network/ heterogeneous network). Those subcategories are determined by the incentive from operators' or users' perspective. Furthermore, we will discuss this basic offloading incentive along with the main technique (delayed offloading). For non-delayed offloading, a key problem is to enhance the network selection mechanism to achieve the best capacity from a comprehensive perspective. For delayed offloading, a key problem is to make a tradeoff between delay and capacity.

Similarly, from operators' and users' perspectives, we will investigate more offloading schemes according to the incentives of cost, energy and rate. WiFi offloading schemes, according to the incentive of cost, make its best to save cost for cellular operators and/or users. A typical problem is the tradeoff between cost, delay, and capacity. The methods may be experimental measurements, analytical model, and development of smart algorithms. The incentive of energy is referred to utilizing various methods for saving energy consumption for cellular operators and/or users. A typical problem is to utilize empirical or theoretical analysis to investigate the tradeoff between energy consumption, capacity, and congestion. In this survey, we comprehensively utilize the concept of rate to describe the performance of average completion time of demanded data transmissions in WiFi offloading. It can determine the quality of services (QoS) directly. In fact, several factors jointly contribute to practical rate such as information transmission rate of a radio access network, switching delay, practical load condition, and congestion.

Moreover, we will identify a very important incentive referred to continuity and extend the common mobile scenario to vehicular scenarios. WiFi offloading scheme usually makes UEs switch radio access network frequently. This may incur disruptions to ongoing communication, damaging users' satisfaction significantly. What is worse, high dynamic of mobile communication environments and fluctuating wireless channels will incur frequent disruptions to ongoing communication, especially for vehicular scenarios. Thus, it is very important to investigate how to alleviate disruptions or maintain continuous communications. Overall, we will investigate various offloading schemes with one incentive after another, presenting their challenges, advantages, limitations, and further directions. We believe that prior surveys which cover more or less the same topic do not have the same depth as provided in this paper. For example, Aijaz *et al.* [7] well distinguished users' perspective from operators' point of view and introduced the status of commercial deployment, but their developed techniques just focused on the evaluation metric of offloading efficiency. More details regarding network selection mechanisms, delayed techniques and other incentives such as cost and energy are not discussed. In addition, Rebecchi *et al.* [8] summarized non-delayed and delayed strategies and also distinguished terminal-to-terminal techniques from AP-based offloading. However, the incentives of the offloading techniques in [8] were not clear while we classify these incentives into five categories. More detailed techniques including cost reduction, energy saving and other emerging techniques (i.e., context-aware offloading, real-time switching, and vehicular offloading) have not been discussed. Thus, our proposed work has a comprehensive classification of WiFi offloading that is different from [7] and [8].

TABLE II
GLOSSARY

Acronym	Full Name
AMUSE	Adaptive-bandwidth Management through USer-Empowerment
ANDSF	Access Network Discovery and Selection Function
AP/APs	Access Point/Access Points
API	Application Program Interface
C2M	Cellular-to-Mesh
CAPEX	Capital Expenditure
CRRM	Cooperative Radio Resource Management
CSP	Content Service Provider
CTP	Cabernet Transport Protocol
D2D	Device to Device
DCF	Distributed Coordination Function
DHCP	Dynamic Host Configuration Protocol
DTN	Delay Tolerant Networking
eMPTCP	energy-aware MPTCP-based content delivery scheme
EPC	Evolved Packet Core
EST	Energy Spectrum Trading
FBMC	Filter Bank Multicarrier
FCC	Federal Communications Commission
FDD	Frequency Division Duplex
GZRP	Genetic Zone Routing Protocol
HHO	Horizontal Handover
HPCM	Heuristic Power Consumption Minimization
HPRO	High PRobability Opportunistic
HSDPA	High-Speed Downlink Packet Access
IETF	Internet Engineering Task Force
IFOM	IP FlOw Mobility
ISP	Internet Service Providers
ITS	Intelligent Transportation Systems
ITU	International Telecommunication Union
LRRM	Local Radio Resource Management
LTEU	LTE-Unlicensed
M2M	Machine to Machine
MADNET	Mobile AD-hoc NETwork
MIG	Mobile Integration Gateway
MIH	Media Independent Handover
MIIS	Media-Independent Information Service
mmWave	Millimeter Wave
MP2MP	MultiPoint-to-MultiPoint
MP2P	MultiPoint-to-Point
MPTCP	Multipath TCP
NFV	Network Function Virtualization
NGSON	Next-Generation Service Overlay Network
NSG	Network Selection Game
NSS	Network Selection Scheme
OFDM	Orthogonal Frequency Division Multiplexing
OPEX	Operation Expense
P2MP	Point-to-MultiPoint
P2P	Peer to Peer
PAPR	Peak to Average Power Ratio
PCMP	Persistent Connection Management Protocol
PSN	Pocket Switched Networks
QoS	Quality of Services
RATs	Radio Access Technologies
RRM	Radio Resource Management
RSS	Received Signal Strength
RSU	Road Side Units
SAS	Spectrum Access System
SCBS	Small Cell Base Station
SDN	Software Defined Networking
SILVIO	Seamless Internet 3G and Opportunistic WLAN Vehicular Internet Connectivity
SINR	Signal to Interference and Noise Ratio
SNR	Signal to Noise Ratio
SON	Service overlay network
TAOS	Traffic Aided Opportunistic Scheduling
TCP	Transmission Control Protocol
TOMP	Traffic Offloading using Movement Predictions
UDP	User Datagram Protocol
UE/UEs	User Equipment / User Equipments
UI	User Interface
UMTS	Universal Mobile Telecommunication Systems
V2I	Vehicle-to-Infrastructure
V2V	Vehicle-to-Vehicle
VANET	Vehicular Ad-hoc Networks
VHO	Vertical Handover
ViFi	V-band WiFi
VoIP	Voice over Internet Protocol
WFP	Windows Filtering Platform

B. Main Contributions

In this survey, we propose a comprehensive guide in the field of WiFi offloading techniques. To the best of our knowledge, this is the first work to classify state-of-the-art WiFi offloading schemes—–combing WiFi offloading techniques with the incentives of WiFi offloading. More specifically, our main contributions are as follows.

1) *Presenting a New WiFi Offloading Classification:* To help researchers quickly find a specific study incentive in their forthcoming research work in this area, we first classify the incentives of WiFi offloading into five categories: capacity, cost, energy, rate, and continuity. Then, we investigate WiFi offloading techniques in terms of each incentive of WiFi offloading. That is, we combine the incentives of WiFi offloading with WiFi offloading techniques. Then we also present their combination characteristics, respectively. We especially investigate the most popular techniques such as delayed offloading, and non-delayed offloading.
2) *Extending Pedestrian Scenarios to Vehicular Applications:* We extend pedestrian scenarios to vehicular applications to meet the incentive of service continuity. After presenting a detailed description of the Drive-thru Internet and related promising protocol issues, we analyze how to deal with the challenges including energy efficiency, connection establishment delays, and frequent disconnections. To specify the potentials of vehicular offloading, we give some

recent advances regarding opportunistic offloading via multi-hop networks.

3) *Pointing Out Potential Research Directions:* To point out the potential approaches to improve the performance of WiFi offloading, we propose ideas for future directions in the area, along with incentives. After we discuss the state-of-the-art trials and results of some experiments, we present detailed comparisons and summarize the characteristics of key technologies for those potential approaches. Finally, we summarize open issues, future research directions, and emerging ideas related to WiFi offloading from organizations in three tables, respectively.

The rest of the paper is organized as follows. In Section II, related issues in heterogeneous networks are investigated, especially the NSS. Section III provides an overview of delayed/non-delayed offloading strategies within the basic incentive of improving capacity. In Section IV, we consider there incentives that can stimulate both users and operators to offload traffic to WiFi, namely cost saving, energy consumption and rate. In Section V, further incentive of continuity is considered to address frequent disruptions and continuous communications. Moreover, Section VI extends the incentive of continuity to vehicular scenarios. Future research directions and challenges are presented in Section VII. In Section VIII, we finally provide the conclusion of this survey. The acronyms are listed in Table II.

II. Related Issues in Heterogeneous Networks

An increasing number of cellular networks and wireless local area network (WLAN) coexist widely, while current mobile devices are capable of accessing both WiFi and cellular networks. To make full use of multiple networks, it is significant to schedule licensed and unlicensed bands well and help the UE integrate its radio resources appropriately. Therefore, the problem regarding the coexistence of WiFi and cellular networks has recently attracted the most attention.

A. Requirements of Coexistence from 5G Wireless Systems

Unlike the previous four generations, high data rate is not the only main feature of 5G. In fact, 5G will show more new characteristics such as lower cost, lower latency, and lower energy consumption compared to previous generations [9]. In addition, ultra densification and high heterogeneity will be essential features of future mobile communication systems. Moreover, unlike traditional cellular communication systems such as LTE macro cell networks, 5G will utilize every air interface such as WiFi and femtocell so as to ensure high QoS and guarantee that users can experience ubiquitous, seamless, inclusive wireless services [10].

Future mobile communication networks should be ubiquitous, heterogeneous and IP-based, thus 5G systems are expected to solve such problems. 5G systems will be converged while integration of different RATs must be implemented [11]. In other words, 5G must not only support the new 5G standard but also be compatible with the previous cellular networks, WiFi and device-to-device (D2D) communications [12]. For that, 5G systems must support D2D communications with the rapid development of the Internet of Things (IoT). The challenge for D2D communication is to provide direct communication, low power, low rate but always online connections [13].

To address the requirements from the 5G systems, an increasing number of research work has been under way. As for reducing the latency to reach the magnitude of 1 ms, there are two main emerging trends which are called cloud-based technologies. The first one is network function virtualization (NFV) while the second is software defined networking (SDN). As for improving the data rates, in addition to MIMO and Millimeter Wave (mmWave) solutions, another popular method is to improve densification to utilize spectrum efficiently per unit area. There are two main methods for high densification. The first is to deploy more base stations, which makes cells shrink to smaller sizes. The challenge in using this method is the increase of the CAPEX and OPEX, which makes this method limited and cannot go far enough. The second method is to utilize multiple RATs in an appropriate way, which leads to heterogeneous wireless networks and coexistence of multiple RATs. However, the first challenge is how to determine the suitable RAT and the best base station or AP that the user should access. The next challenge is to support continuity and avoid interruptions in switching progress between heterogeneous wireless networks. The third challenge is how to offload data from cellular networks to WiFi and/or other RATs efficiently. The fourth challenge is the coordination and the management of multiple networks, while the self-organization between heterogeneous networks is a promising way. The last challenge is to support high mobility between heterogeneous wireless networks without damaging the users' experience. These challenges regarding switching and offloading in heterogeneous networks will be discussed later.

B. Network Selection Scheme

No single network radio access technology can simultaneously provide low cost, low energy consumption, low latency, high bandwidth and high throughput data services to a large number of mobile users. Network selection on the mobile terminal is proposed to make a decision and switch wireless interfaces in heterogeneous networks, and such a decision mechanism can be initiated by smart terminals or wireless systems according to the status of these interfaces.

1) Factors in Network Selections: A NSS allows a UE to select the optimal access network from a heterogeneous network including a WiFi network and a cellular network. There are different options and algorithms to choose the best RAT between different networks, but not the same factors are of concern. In order to determine the best RAT, the traditional consideration is Received Signal Strength (RSS), while the data rate and congestion is ignored. In general, the considerations also include signal to interference and noise ratio (SINR), the instantaneous load, user's preference, etc. As shown in Table III, there are three kinds of factors that NSS may

TABLE III
CONSIDERATIONS FOR NSS

Category	Considerations
Device Status	Application type, Battery life, Physical speed, Network selection history
Channel Status	Radio channel quality, Physical obstructions, Relative position
System Status	Energy efficiency, Average throughput, Response time, QoS, Port blocking, Backhaul capacity

take into consideration, including the conditions of the terminals, the channel conditions, and the network performance of multiple wireless systems.

A selection algorithm for network-cooperation-based radio access technology has been proposed in [14], where the RAT algorithm utilizes suitability to make the best choice between WiFi and high-speed downlink packet access (HSDPA). It is a potential approach to provide gain by assigning the user terminal to the optimal network based on the communication type and the network load.

2) Handover Strategy: The vertical handover (VHO) is a key method to optimize the utilization of radio resources between different wireless access networks. The traditional handover based on RSS is suitable for horizontal handover (HHO), but is not suitable for making a VHO decision. Thus, many research attempts explored utilizing the characteristics of heterogeneous wireless networks. The work in [15] shows that VHO based on SINR has a higher throughput for terminal and network compared to RSS-based VHO. Considering different factors with a cost function, the policy-enabled VHO algorithm was presented in [16]. As an enhanced algorithm, a multi-criteria VHO algorithm was presented in [17]. A VHO algorithm considering the bandwidth and RSS as the key factors was presented in [18]. The shortcomings of these algorithms are that they may cause ping-pong effect if a tradeoff between different performance metrics is not considered. Moreover, in order to select a best access network, further investigation is needed to avoid switching to the network with bad performance and eliminate the ping-pong effect in the case of coexistence of LTE and WiFi networks.

In order to execute the network selection and allow it to reach its best performance, it is potential to utilize a cost function which provides flexibility in selecting decisions so as to balance the different factors. In fact, a cost function can be considered as a sub-module of VHO. It focuses on seamless handover that can improve energy efficiency of mobile devices. A selection scheme based on the cost function is a flexible algorithm since relevant factors can be considered to make a tradeoff. Few researchers utilize 802.21 media independent information service (MIIS) which records available networks and parameters to provide additional information. This will allow devices to retrieve information regarding different access networks and establish a net map. For each available access network, such additional information is critical to the handover procedure. Simulations in [19] show that it selected the optimal network compared to the conventional trigger mechanism. The switching is triggered at an appropriate time in order to enhance the connectivity. In addition, this solution optimizes the energy consumption of multi-radio devices significantly. It is able to enhance continuity of ongoing sessions and support seamless handoff in this intelligent algorithm. However, some mobile devices may not support MIIS, and further local optimization approaches are needed to save scanning energy consumption. It is extremely important to explore novel motion detection schemes, such as algorithm utilizing the parameters collected from vehicles' accelerators.

3) Access Network Selection Algorithms: In order to help UEs discover non-cellular wireless access networks in their vicinity, 3GPP has specified handover algorithms for the access network discovery and selection function (ANDSF) to control the handover operations between 3GPP and non-3GPP networks [22], [23]. The main function of ANDSF is to detect the available access networks in the proximity based on current geographical information collected from UEs. ANDSF is also responsible for managing the available access network by prioritizing them based on operators' policies. It is a server-based solution and provides UE with a prioritized list of non-3GPP wireless access networks based on predefined policies when the UE requests ANDSF information from the ANDSF server. So, the UE should provide its location information or cell ID before sending any requests to the server. ANDSF is deployed in evolved packet core (EPC) by operators. Thus, a UE has to establish the connection to EPC through 3GPP access networks to infer whether any non-3GPP access networks, such as WiFi, in the vicinity are available or not.

Previous research shows that the load controlling of WiFi offloading could be well achieved by setting the signal strength threshold and the network distance threshold appropriately. In fact, as far as the signal to noise ratio (SNR) threshold is concerned, ANDSF model is classified into different schemes according to different thresholds which can be set to 0 or a fixed value based on the network condition. As shown in Table IV, several ANDSF schemes based on different SNR thresholds are compared. In algorithm 1, by setting the SNR threshold SNR_{min} to 0, offloading is initiated when a WiFi access network is detected, which is on-the-pot offloading. In algorithm 2, $SNR^{1}_{min}(load^{1}_{i})$ indicates the way to set the value of SNR^{1}_{min} in an implementation, where SNR^{1}_{min} denotes the minimum SNR value of WiFi access network 1 for offloading received by one user i. $SNR^{1}_{min}(load^{1}_{i})$ is a function of $load^{1}_{i}$ which represents the load of WiFi network 1. Further, these parameters should be set according to the following policy: SNR^{1}_{min} increases with $load^{1}_{i}$, while $load^{1}_{i}$ increases with $load_{system}$ so as to balance the load of heterogeneous networks, where $load_{system}$ denotes the actual load of the hybrid system. In a balanced mechanism, $load_{system}$ will be maintained below C_{system}. As for the result of algorithm 2, the equation $access_{i} = WiFi$ indicates that UE i is bound to select WiFi 1 to offload when the SNR^{1}_{i} of WiFi network 1 with the best SNR received by UE i equals or exceeds the SNR^{1}_{min} of WiFi 1 network. In a nutshell, these parameters of algorithm 2 should be set according to the principle that the WiFi network with low load is more likely to offload mobile data traffic.

TABLE IV
FUNCTIONS AND CONDITIONS OF FIVE ALGORITHMS

Algorithm	Function and Condition	Input Parameter	Variables Set
1. On-the-spot Offloading [20][53]	$Access_i = \begin{cases} WiFi & SNR_i^1 \geqslant 0 \\ LTE & otherwise \end{cases}$	SNR_i^1	0
2. Fixed SNR Threshold [21]	$Access_i = \begin{cases} WiFi & SNR_i^1 \geqslant SNR_{min}^1(load_i^1) \\ LTE & otherwise \end{cases}$	SNR_i^1, $load_i^1$	SNR_{min}^1
3. ANDSF Model based on Cell-ID [22]	$Access_i = \begin{cases} WiFi & SNR_i^1 > 0 \\ LTE & otherwise \end{cases}$	SNR_i^1	SNR_{min}^1
4. ANDSF Model based on Position [22]	$Access_i = \begin{cases} WiFi & Distance_i < D_{thr} \\ LTE & otherwise \end{cases}$	$Distance_i$	D_{thr}, SNR_{min}^1
5. Hybrid ANDSF Model [22]	$Access_i = \begin{cases} WiFi & \begin{cases} Distance_i < D_{thr} \\ SNR_i^1 > 0 \end{cases} \\ LTE & otherwise \end{cases}$	$Distance_i$, SNR_i^1, $load_i^1$	D_{thr}, SNR_{min}^1

In algorithm 3, WiFi APs are discovered if they are located in the region of UE's current macro cell ID, then AP 1 with the strongest SNR_i^1 received by UE i will be selected. In algorithm 4, WiFi APs are discovered if they are located close enough to UE i within the distance threshold D_{thr}, then AP 1 with the strongest SNR_i^1 received by UE i will be selected. In algorithm 5, AP 1 is selected if it meets the conditions of algorithm 3 and algorithm 4. It is noteworthy that algorithm 3, algorithm 4, and algorithm 5 have an adjustable parameter *LTE ISD* (LTE Inter Site Distance) which represents the distance between two BSs [24]. In fact, the performance of LTE networks is determined by *LTE ISD* on average. In other words, the smaller the *LTE ISD*, the higher the bit rate LTE offers than that of the WiFi on average, the fewer users will choose the WiFi offloading. The larger the *LTE ISD*, the smaller the bit rate LTE offers than that of WiFi on average, the more users will use WiFi offloading. Thus, the smaller the *LTE ISD*, the smaller the effect of WLAN on users, which means that a user can keep a high bit rate and throughput no matter D_{thr} or *SNR* is small or large in terms of algorithm 5. Algorithm 5 shows its advantage when *LTE ISD* is fairly large. In fact, in algorithms 3, 4, and 5, considering the effect of *LTE ISD* is realized by setting D_{thr} appropriately that represents the discovery distance threshold of the LTE network. The policy is that both the values of D_{thr} and $SNR_{\min}$ vary with the actual value of *LTE ISD* in the implementation.

Overall, whether the mobile traffic is offloaded to a WiFi network or not is determined by the value of the $SNR_{\min}$ in terms of WiFi access networks and the value of D_{thr} in terms of LTE access networks. The higher the $SNR_{\min}$, the higher the offloading condition is, the fewer users will offload the mobile traffic to the WiFi. In detail, algorithm 1 is the best algorithm when the quality of the WiFi system is high, which is the minimum condition of WiFi offloading that encourages as more users as possible to migrate their traffic. On the contrary, the higher the load of WiFi 1 ($load_i^1$), the higher the value of $SNR_{\min}$, the higher the condition of WiFi offloading will be, the fewer the users will offload their traffic. In addition, the smaller the D_{thr} is set, the fewer the users that will offload their traffic to the WiFi networks. Ultimately, whether a UE offload its traffic is determined by the simultaneous performance of the WiFi networks and the LTE networks.

Access selection mechanism is able to guide users to make use of both 3GPP and non-3GPP radio resources. As the capacity of the multi-access system is larger than the capacity of a single cellular access system, these algorithms are capable of helping operators increase capacity and bit rate of the total network significantly. The goal of ANDSF is to increase the bit rate, therefore the network selection decisions must take two factors into account that include signal quality and system load. The 3GPP only allows selecting the mechanism with the current priority, which means that the highest priority network should be selected. In fact, this priority selection mechanism is actually not the best for heterogeneous wireless networks. For example, this mechanism does not consider the real-time performance of the network. Moving devices are prone to losing the signal because of buildings or other obstructions which generate damage and attenuation to a radio signal. Current solutions are to make devices connect to the WiFi with the highest signal strength possible. In other words, offloading is initiated when the maximum value of *SNR* of available WiFi networks matches the condition: $SNR > SNR_{\min}$. Setting $SNR_{\min}$ is essential that $SNR_{\min}$ must be set on the basis of the quality of the signal and load of the system. The future challenge is to provide WLANs with the ability of load analysis and mission control in order to ensure that an overload will not occur in a WLAN network. Actually, other propagation conditions also contribute to SNR_{min} and further investigations are needed to survey the different values of SNR_{min} within different *LTE ISD*. Moreover, further research is needed to extend the scenario to multiple users by taking interferences between multiple users into considerations, adopting SINR instead of SNR.

However, in order to infer the decision based on information collected from users, ANDSF may lead to unnecessary scanning on UE while the energy efficiency may decrease. A promising approach is to avoid scanning of WiFi access networks and UE connections unless it is stationary since offloading is not efficient for UEs with high mobility. To probe the motion state of UEs and turn on their WiFi interfaces automatically when they are stationary, Kim *et al.* [25] proposed a novel user high-level motion detection approach, and evaluation showed that unnecessary scanning significantly reduces the energy consumption compared to the Android WiFi connection manager that utilizes periodic scanning. Furthermore, Hagos and Kapitza [24] argue that the probability of hotspots can also affect the propagation conditions and the system load. However, further research is needed to enhance the energy efficiency on UE with mobility to prolong the battery life of mobile devices, such as prediction schemes based on historical information, geographical information and statistical data.

Huang *et al.* [26] developed the first empirically derived comprehensive power model of a commercial LTE network. They used real traces of users and found that the power efficiency of LTE networks is one twenty-third of that of WiFi networks. Thus, to investigate the handover mechanism between WiFi networks is significantly needed for energy savings. It is notable that a significant amount of energy will be consumed during scanning. Doppler *et al.* [27] proposed a novel light-scanning assistance combined with ANDSF to limit the amount of scannings during the period when the UEs are not moving. In this model, the ANDSF server provides UEs with not only the number of accessible APs but also the channel knowledge (5 GHz band/US TV WS band). This information can well help the UE select the scan mechanism (passive scanning/active scanning). More specifically, the UE chooses passive scanning if the number of identifiers of APs exceeds 55 and chooses active scanning otherwise. Evaluation results show that the energy consumption quickly drops with higher AP density. The energy consumption also reduces when channels are only scanned and the UE moves over 5 meters. However, this model does not provide UEs with up-to-date knowledge about the channels and just assumes that the UE knows the existing profile. The up-to-date knowledge will increase the complexity of the ANDSF server. Hence, further work is still needed to investigate its performance in a real scenario.

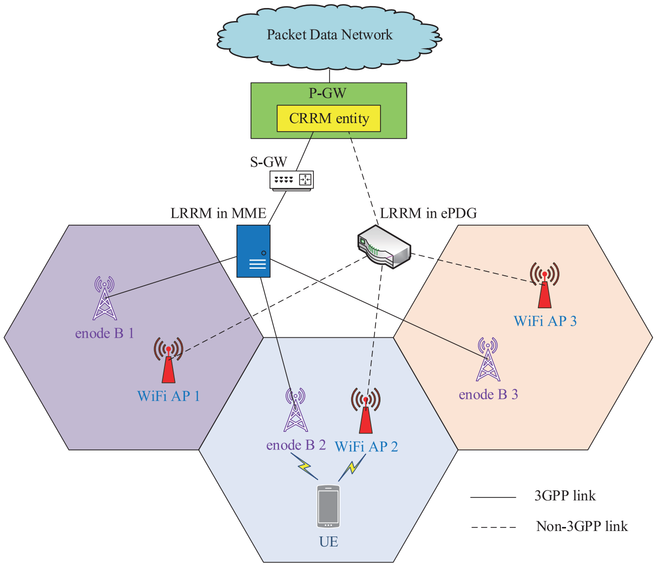


Fig. 2. Cooperation achieved by coordinated radio resource management. In an evolved packet core net, CRRM entity is installed in the existing module P-GW, while the LRRM is integrated into the existing module MME or ePDG. In fact, as the location of CRRM is of concern, the scenario in the figure is kind of a centralized algorithm.

TABLE V
CRRM ALGORITHMS CLASSIFICATION

Differences **CRRM** **Items**	**Centralized [31]**	**Distributed [32]**
CRRM Entity position	in the core network	in UE
System policy	considered well	not guaranteed
Users' QoS	not guaranteed	considered well

C. Cooperative and Management

In addition to network selection and discovery schemes, it is vital to enhance the offloading scheme by scheduling its multiple networks in a better way such as cellular-network-assisted cooperation. To achieve cooperation between multiple networks, a popular way, namely "the master/solution" presented in [28] is to utilize the cellular network as the always-connected access network to control switching while other access networks are utilized to offload traffic by opportunistic routing. However, it takes a lot of expenditures for the master/solution to control multiple networks completely. In addition, intentional networking was proposed in [29] to provide hints to the system in order to select the optimal wireless interface. Those approaches utilize cellular networks as their masters in multiple RAT networks.

In the recent literature, cooperation between networks is commonly achieved by cooperative radio resource management (CRRM) as shown in Fig. 2. Traditional local radio resource management (LRRM) is suitable for homogeneous wireless networks, but it is not able to support the management between heterogeneous networks. CRRM was proposed to allocate UEs to the optimal interface in heterogeneous wireless networks by managing radio resource utilization appropriately, which is a crucial way to improve capacity. In general, the aim of CRRM is to improve the utilization efficiency of radio resources, enhance user's satisfactions, and increase operator's profit. It allows mobile devices to communicate across several networks directly based on a cross-system information, responding to the UEs specific requirements. In fact, many enhanced offloading schemes are based on the cooperation between licensed and unlicensed bands, and further details are discussed in the next section. The need of CRRM for the next generation wireless networks was discussed in [30], and different approaches for the distribution of LRRM and CRRM entities were also compared.

There are three modules in a typical architecture of CRRM, and the first one is responsible for collecting user-specific information. The second is to handle the information in accordance with the specific requirements of UE, and the third is to initiate a new switch based on load balancing conditions and status. Some scholars have conducted simulations of service model based on load adaptive delay constraints. Evaluations in [2] provided that the proposed CRRM performs well on the radio resource management in the heterogeneous wireless networks. This process utilized a protocol based on adaptive RAT to support cooperation between WiFi and cellular networks. This work discovered the optimal load threshold to maximize the total throughput of cooperation between two kinds of networks.

As shown in Table V, many CRRM algorithms can be categorized into centralized algorithms [31] and distributed algorithms [32]. In a centralized CRRM algorithm, CRRM entity is deployed in the core network as shown in Fig. 2. The CRRM entity is deployed within UEs in a distributed CRRM algorithm. The centralized algorithm considers the system policy well but cannot guarantee user's QoS requirements. The distributed algorithm considers the user's preference but it does not consider the system status and policy well. The distributed algorithm outperforms the centralized algorithm in terms of dynamic radio resource control. The Distributed CRRM allows users to select the best interface and improve their satisfaction. Another benefit of distributed CRRM is that there is no need to change the existing infrastructure for cellular network operators. However, it is promising that a hybrid RAT selection algorithm that utilizes the network to assist distributed devices in making a decision by providing information and policies so as to select the most efficient RAT. Mihovska *et al.* [33] presented various architectures of CRRM and provided several kinds of algorithms and requirements for wireless interfaces. In order to support user's mobility and fast decision of switching, they proposed a novel radio resource management (RRM) framework for an IMT-Advanced scenario by network-controlled policy mechanism, and combined centralized and distributed RRM. In this framework, the controlling functions are established in the core net. In fact, IEEE P1900.4 Protocol was proposed in [34] to support this hybrid cooperation, and more details were presented in [35]. However, further investigations are needed to optimize and evaluate the intelligent hybrid CRRM approach in different scenarios.

To evaluate the performance of CRRM, Tolli *et al.* [36] studied the benefits for both real-time and non-real-time traffic by relatively simple Matlab simulations in terms of load balancing and capacity. They demonstrated that a slight increase of capacity can be achieved in terms of multimedia service, while a significant improvement of interactive capacity can be achieved by CRRM. Based on the results, they argued that CRRM was most important for interactive and streaming connections with high bit rates. However, no delays are assumed in their simulations, which needs further practical considerations.

III. Basic Incentive With Improving Capacity

To address the data explosion and enhance network capacity, a significant method in the heterogeneous network is WiFi offloading. Given that there is already a widespread deployment of WiFi networks, WiFi offloading might be the most practical method, in contrast to other methods such as femtocell. As shown in Table VI, the differences between WiFi offloading and femtocell offloading are provided. According to whether WiFi offloading scheme defers the data traffic to WiFi networks, current offloading schemes are classified into two basic classes: Non-delayed offloading and Delayed offloading. This is the most typical method of classifying. Almost all the recent offloading schemes can be seen as extensions to these two basic offloading schemes.

TABLE VI
Differences Between WiFi Offloading and Femtocell Offloading

Offloading / Differences	WiFi offloading	Femtocell offloading
Frequency band	Unlicensed band	Licensed band
Standard	802.11 standard family	3GPP
Deployer	Operators/The third party	Subscribers

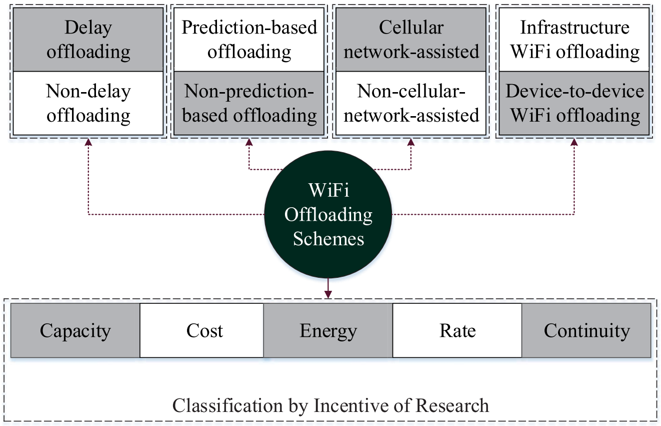

Delay offloading
Non-delay offloading
Prediction-based offloading
Non-prediction-based offloading
Cellular network-assisted
Non-cellular-network-assisted
Infrastructure WiFi offloading
Device-to-device WiFi offloading
WiFi Offloading Schemes
Capacity
Cost
Energy
Rate
Continuity
Classification by Incentive of Research

Fig. 3. Current WiFi offloading schemes can be classified into different categories according to different metrics.

Actually, there are other methods for classifying according to different metrics as shown in Fig. 3. First, according to the incentive, offloading schemes can be classified into five categories: *Capacity*-centric, *Cost*-centric, *Energy*-centric, *Rate*-centric, and *Continuity*-centric. Second, according to the control mechanism, offloading schemes can be classified into two categories: cellular-network-assisted and non-cellular-network-assisted offloading. Third, offloading schemes can also be classified into two categories: infrastructure WiFi offloading and WiFi Direct offloading, according to whether the offloading paths are based on infrastructure WiFi or WiFi direct. Last, offloading can be further classified into two types: prediction-based offloading and non-prediction-offloading, according to the modeling method. In short, whether WiFi offloading scheme defers the data traffic to WiFi networks is the basic method of classification.

Regardless of non-delayed offloading and delayed offloading, traditional WiFi offloading schemes are devised to achieve a better performance of capacity. Many studies exploit various methods to measure and evaluate the performance of WiFi offloading schemes with the metric of capacity. Furthermore, capacity can be further classified into two subcategories: amount of the total mobile data traffic being offloaded and the throughput per user (of LTE network/ heterogeneous network). It is notable that the performance metric of the amount of the total mobile data traffic can be further specified by the concept of offloading efficiency [37] or offloading ratio [41]. Offloading efficiency is defined as the ratio of the amount of offloaded data to the total amount of data. A detailed classification is given in Table VII, considering the incentive of improving capacity.

TABLE VII
TRADITIONAL OFFLOADING ACCORDING TO THE BASIC INCENTIVE OF ENHANCING CAPACITY

Incentive / **Subject**	**Capacity**	
	Amount of the total mobile data traffic being offloaded	Throughput per user (of LTE network/ Heterogeneous network)
Non-delayed offloading	Network Selection Scheme	Network Selection Scheme
	[38][39]	[44][24][25][45]
	Deploying Strategy of WiFi Network	Deploying Strategy of WiFi Network
	[40][41][42][43]	[46]
		Optimization Problem in Modeling
		[47][46][48][49][50]
Delayed offloading	Tradeoff between Delay and Amount	Tradeoff between Delay and Throughput
	[51]	[55]
	Users' Trace/Real-Time-Measurement-based	
	[52][53]	
	Network Selection Scheme	
	[54]	

A. Classification Considering Delay Technology

1) Non-Delayed Offloading: Current smart-phone platforms have offered the simplest offloading mechanism: on-the-spot offloading. This scheme is based on WiFi first algorithm where users connect to a WiFi access network whenever WiFi coverage is available [40]. Then UEs shift their data traffic to a WiFi network. In other words, UEs offload their data traffic only if the received signal strength of WiFi access is non-zero, otherwise they use cellular networks to transfer the data traffic. This offloading scheme uses the network selection algorithm based on RSS. The key is to evaluate the RSS and to model the propagation path loss of different RATs. In fact, International Telecommunication Union (ITU) presents guidance to model LTE propagation [56] and WiFi propagation [57]. Hu *et al.* took account of building height in their propagation path loss modeling [58]. Inspired from this work, Thiagarajah *et al.* performed a 3D simulation of the coverage of the LTE base stations and WiFi AP nodes, considering the actual building height [44]. They use the easiest selection algorithm "WiFi First" by the client based on RSS, as a kind of NSS. This simulation indicates that around 50% of the users are offloaded to WiFi, while the total capacity usage of LTE networks can be only reduced by 2.75%. On the other hand, the average capacity experienced by the LTE users increased by nearly 100%.

Nevertheless, on-the-spot offloading is still the simplest offloading mechanism as a non-delayed offloading scheme. It shows many limitations and needs to be developed into a smarter mechanism. First, in view of the practical density of WiFi AP nodes, on-the-spot offloading is only suitable for urban areas. Second, it takes no account of the characteristics between delay-sensitive and delay-tolerant application data traffic. Third, it does not take into account the best portion of traffic being offloaded for different application data. Fourth, it does not consider the instantaneous network conditions or historical information related to wireless access networks. Finally, the access network selection algorithm needs to be more smart and takes more information into account. Thus, more efforts to exploit different technologies to augment non-delayed offloading scheme are needed.

2) Delayed WiFi Offloading: Note that many applications suffer transfer delays without significantly damaging the functionality of these services, cellular traffic can be offloaded to WiFi when users are willing to delay their traffic. The reason lies in that the WiFi network may not be in range anywhere. So, waiting to connect to a WiFi network will cause a WiFi offloading delay. These tolerant data traffic that can be delayed may include movies and software downloads. A mobile device can upload or download data when it is in the range of WiFi networks but there is a pre-set time deadline for users to wait. In other words, UEs switch to a cellular network to transfer data when the delay is larger than the deadline.

It is notable that WiFi offloading delay is appropriate for a scenario where the density of WiFi deployment is considerably low and the continuous data links between mobile nodes and services via WiFi networks are usually interrupted when users move away from the range of the current WiFi access to another one. In addition, bulk data is particularly suitable for delayed WiFi offloading as far as bandwidth, energy and cost are concerned [59]. It is worth pointing out that some current applications on smart-phone platforms are a subclass of delayed offloading schemes. In fact, delayed offloading is relatively new to current smart-phone platforms, in contrast to on-the-spot offloading which is prevalent in smart-phone platforms. In addition, the concept of Delay Tolerant Networking (DTN) is close to delayed offloading. DTN can be deployed by operators to transfer bulk data via the WiFi network path across the Internet rather than the cellular network path since WiFi offloading is economically beneficial for both users and operators.

Delayed offloading has already attracted many focus of researchers to evaluate the WiFi offloading efficiency. Lee *et al.* conducted a quantitative study on the performance of WiFi offloading by experimental analysis of metropolitan areas in South Korea [53]. They used an iPhone application that can track WiFi connectivity and recruited 97 users to collect data during a period of about two weeks. They used a whole-day statistics and trace-driven simulation to measure the efficiency of two offloading schemes. For the on-the-spot offloading, this study indicates that WiFi can offload about 65% of the total cellular data traffic without any delay. However, substantial

TABLE VIII
OFFLOADING EFFICIENCY AND ENERGY SAVING IN SIMULATION [53]

Offloading Scheme	**On-the-spot**	**100 s delay**	**1 h delay**
Offloading Efficiency/Gains	65%	+(2% – 3%)	+29%
Energy Saving/Gains	55%	+3%	+20%

gains in the amount of the total offloading traffic can be achieved when they use the delayed offloading with setting the delay deadline fairly larger than tens of minutes. In more detail, as shown in Table VIII, the achievable gain over on-the-spot offloading is less than only 2%–3% with 100-s delays because of long interconnection times. Moreover, the achievable gain over on-the-spot offloading can be beyond 29% with a deadline of one hour and longer. The average completion time of data transfers is shorter than their deadlines. In addition to offloading efficiency, this study also indicates that on-the-spot offloading can achieve about 55% energy saving without any delay. The delayed offloading can achieve 3% energy saving gain over on-the-spot offloading with 100 seconds deadline, and achieve 20% energy saving gain over on-the-spot offloading with one hour deadline. In fact, for this study, they focused on the uplink scenarios, and further efforts are needed to extend it to downlink cases.

Nevertheless, it is not well investigated in the literature in terms of figuring out the expected amount of data that can be offloaded under various conditions by analytical models. Various conditions may include WiFi data rates and residence time of wireless access networks. Suh *et al.* developed an analytical model on offloading efficiency for on-the-spot WiFi offloading and delayed WiFi offloading [37], in contrast to the experimental analysis discussed in [53]. This is the first analytical study on the relationship between the amount of data that can be offloaded and various environments. It demonstrates that delayed offloading can improve the WiFi offloading efficiency of non-delayed offloading by 179-198%, 234-296%, and 319-489% when the deadline is 10%, 20%, and 40% of the average WiFi residence time, respectively. Moreover, it also reveals how the temporal coverage of WiFi networks, the average data rate of WiFi APs, and the average session duration affect the WiFi offloading efficiency. However, further work is needed to trade off between the delay deadline and the amount, taking account of users' satisfaction.

B. Basic Incentive of Improving Capacity

In addition to efforts focusing on evaluating WiFi offloading efficiency, an increasing number of attempts exploit different technologies to augment offloading schemes and improve capacity for non-delayed offloading and delayed offloading. As shown in Table VII, the incentive of improving capacity can be classified into two subcategories: the amount of the total mobile data traffic being offloaded and the throughput per user (of LTE network/ heterogeneous network). In fact, this classification is determined from the perspective of the operator or user. To the best of our knowledge, this table is the most comprehensive classification considering the incentive of improving capacity in the literature.

1) Non-Delayed Offloading: [38]–[50].

a) For total amount: To improve the total amount of data being offloaded, Kou *et al.* [38] proposed an offloading algorithm deployed in mobile integration gateway (MIG). This algorithm is a kind of NSS and is based on the channel quality. Malandrino *et al.* [39] proposed a dynamic offloading scheme. In this scheme, cellular operators utilize a policy server to capture the user behaviour and refine the current policy if needed. The policy is actually a developed network selection scheme based on ANDSF, as specified in 3GPP R12 [60]. This work was a first attempt to exploit existing 3GPP standards for offloading strategies to augmenting WiFi offloading scheme and investigate offloading efficiency.

In addition, Bulut and Szymanski [41] focused on the problem of WiFi AP deployment for efficient offloading. They analyzed a large scale real user mobility traces and proposed a deployment algorithm of WiFi AP in a metropolitan area. They proposed to deploy the APs to the location with the highest density of user data access request. They also find the optimal deployment by formulating the problem as an Inter Linear Programming problem [61], solving it using the IBM ILOG CPLEX software package [62]. Simulation results demonstrate that this algorithm can achieve higher efficiency than that of previous works. However, it does take into account more information such as the size of data traffic and bandwidth. Oliveria and Carneiro [42] took into account mobile users' context and content, such as their trajectories, scenario interactions, and traffic demands. They leverage the restriction imposed by transportation modes to capture users' content. This work employs a realistic traffic model to provide its higher offloading ratio than the current approach in the literature. Further work is needed to investigate the impact of changes on the mobility.

b) For throughput per user: To improve the throughput per user, there were some previous attempts to focus on network selection schemes to augmenting WiFi offloading. Thiagarajah *et al.* leveraged the easiest selection algorithm "WiFi First" by client based on RSS, where "WiFi First" is close to the idea of "On-the-spot" offloading. They performed simulation on the basis of LTE propagation model, WiFi propagation model, and building-contour height modeling. Hagos and Kapitza [24] proposed an optimized SNR-threshold-based handover solution as an extension to 3GPP standard for ANDSF framework [22], [23] for WiFi offloading. However, ANDSF is a client-based scheme that does not take into account the energy consumption of scanning. To avoid unnecessary WiFi scanning and connections, Kim *et al.* [25] proposed a novel ANDSF-assisted WiFi control method based on the user's motion state, such as walking, driving, and stationary. It requires the operator to deploy an ANDSF server to interact with users, performing an automatic WiFi control mechanism. It is notable that this control mechanism requires users to establish a cellular network connection to transmit ANDSF information in advance.

To ensure the throughput per user, Kim *et al.* [46] investigated how many WiFi APs are needed with a proper number of users per WiFi AP, subjective to limitations of CAPEX/OPEX. They propose an analytical model on average per-user WiFi throughput. Then they set the target value of this model so as to find the minimum required number of WiFi APs in a heterogeneous network, solving it within an optimization problem. However, it does not consider any information to capture the differences between different areas. Further work is needed to investigate how spatial and historical statics affect the minimum required number of WiFi APs, considering users' trajectories and frequency of data access requests.

To investigate the per-user throughput analytically, Garcia *et al.* [48] proposed a very simple analytical model to calculate the per-user throughput on the uplink and downlink. Extensive real-world measurement campaigns indicate that this model can successfully match actual results. It can assist cellular system engineers in predicting the offloading potential. To maximize the per-user throughput in a heterogeneous network, Jung *et al.* [47] proposed a novel network-assisted WiFi offloading model. This model collects network information, such as the number of users in WiFi network and their traffic load. They solve the model as a maximization problem to achieve the maximum per-user throughput, inferring the specific portion of traffic should be offloaded for users. Then users are required to offload their traffic with the probability of the specific portion's value. However, this work does not take into account the difference of voice and data services. Roy and Karandikar [49] investigated how to maximize the per-user throughput by formulating this problem as an incentive function, considering a constraint on the voice blocking probability. In addition, Shoukry *et al.* [50] proposed a novel approach that prefetches multimedia content to users before requesting to ease congestion. This proactive scheduling framework is developed by formulating the scheduling problem with a stochastic model for the user behavior processes of interest to maximize the cached throughput.

2) Delayed WiFi Offloading: [51]–[55].

a) For total amount: To take into account the users' satisfaction, a tradeoff between the offloading amount and the delay is needed for the offloading scheme. Cheng *et al.* [51] presented an analytical relation between the offloading effectiveness and the average service delay. The average service delay is defined as the average time the data services can be deferred for WiFi availability. They established a queueing model and mobility model in a vehicular environment to avoid a longer delay than the expected service delay. Simulation results validate their analysis and indicate the relationship between offloading effectiveness and average service delay. However, this work only uses the infrastructure WiFi access networks without considering the continuity problem. A planned deployment of WiFi AP may help augmenting WiFi offloading performance.

Actually, WiFi networks may show poor availability and low throughput in practical metropolitan areas. Balasubramanian *et al.* conducted a joint offloading system called Wiffler to overcome the poor performance [52]. For delay tolerant applications, Wiffler utilizes a *WiFi connectivity prediction model* to defer application data on WiFi (delayed offloading). The completing time for the delayed offloading is limited to a threshold. Otherwise, Wiffler transfers the application data on cellular networks. For delay sensitive application, Wiffler utilizes a *fast switching* technology to quickly switch to a cellular network when WiFi is unavailable (on-the-spot offloading). Thus, Wiffler is a developed integration of on-the-spot offloading and delayed offloading. This study shows that Wiffler can significantly reduce the load of cellular networks due to delayed offloading. In more detail, WiFi can offload 10%–30% of the total cellular data traffic without any delay. Moreover, WiFi can offload 45% of the total cellular data traffic for a 60 second delay tolerance. However, these results are based on the traces from transit buses or war-driving, which might incur frequent disconnects. It can not account for the practical scenarios in users' normal daily lives. Lee *et al.* [53] performed the first trace-driven simulation using the acquired whole-day traces in South Korea. It indicates delayed offloading with 100 seconds delay can achieve only 2%–3% additional gains over on-the-spot offloading, in contrast to the work in [52]. Furthermore, simulation results show that their work can be used to predict the average performance of offloading for a given WiFi deployment condition with about 10% margin error.

To associate users with delayed traffic in a nearby offloading zone, Mohamed *et al.* [54] proposed a network-based adaptive scheme for offloading zone association. This scheme aims to maximize the total offloading amount. A novel network-based online algorithm is proposed to select the optimal offloading zone for a user with delayed traffic, as an enhanced ANDSF for delayed offloading. They present a cloud cooperated heterogeneous network to control the delayed offloading process in a centralized manner, where the offloading zones and the cellular BS are linked to the centralized radio access network (C-RAN). Simulation shows that it can outperform the traditional ANDSF for delayed offloading in terms of tradeoff between the total offloading amount and the delay.

b) For throughput per user: To investigate how long a user should wait for the offloading and how much data should be offloaded, Zhang and Yeo [55] proposed a utility function to quantify the tradeoff between the offloaded volume and the QoS. This utility function takes into account four parameters: the average offloaded volume during the delay period *t*, the size of the requested content, the maximum delay tolerance of the user with respect to the requested content, and the satisfaction function. They use a contact-sequence enumeration and semi-Markov model to predict the average data volume offloaded during a future period. Then they translate the tradeoff problem into finding the optimal delay bound that maximizes the utility function, achieving the best tradeoff. This work can help users determine the optimal handing-back time when users stop waiting for WiFi offloading. Further work is needed to take into account more information such as content type, users' mobility, and deployment of WiFi APs.

IV. Classification According to Incentive: Cost, Energy, and Rate

To offer a guidance for the future study, the state-of-the-art on WiFi offloading in the literature categories into four subcategories: cost, energy, rate, and continuity, according

TABLE IX
CLASSIFICATION ACCORDING TO THREE INCENTIVES

Incentive / Subject	Cost	Energy	Rate
Operator	Based on Auction Mechanism	Modeling & Optimization Problem	Traffic Steering policies for Load Balance
	[63][64][65][66][67][68]	[79][80][81]	[82][98][99]
	Modeling & Optimization Problem	Radio Resource Management	Optimization Problem for Load Coupling
	[69][70][63][71][66][73][74][68][75][76]	[82]	[100][101]
	Leasing the Third Party APs	Leasing the Third Party AP	Congestion Aware NSS
	[63][70][66][67][68][76]	[79]	[6]
User	Modeling & Optimization Problem	Measurement-based Analysis	Modeling & Optimization Problem
	[72][77][78]	[83][84][85][86][87]	[102][103]
	Cache Coordination Algorithm	Algorithm & Protocol	
	[78]	[88][89][84][90][86][91][92]	
		Modeling	
		[83][93][94][95][87][92]	
		Network Selection Scheme	
		[83][96]	
		IP Flow Mobility & Dividing Flow into Subflows	
		[85][91][95][97]	

to different incentives. These different incentives are also illustrated in Fig. 3 (Section III). Unlike the capacity mentioned above, cost, energy and rate are regarded as particular incentives that are investigated to stimulate operators and users to accept WiFi offloading. Detailed classification according to three incentives (cost, energy, and rate) are shown in Table IX. In addition, classification according to continuity will be further discussed in the next section.

A. Cost

Note that the high potential of WiFi offloading may not make it accepted or adopted by both the users and cellular network operators. If users do not benefit from the offloading scheme, they may not be willing to defer their traffic to WiFi since they can transmit traffic instantly by cellular networks. On the other hand, the operators have to invest a lot to deploy WiFi APs in a heterogeneous network. WiFi offloading will be well accepted and deployed in the heterogeneous network if WiFi offloading is economically beneficial for both the operators and users. Thus, it is necessary to investigate how the operators and users can benefit from WiFi offloading.

1) Operators' Perspective: In order to increase the capacity of cellular networks and deal with the demand of traffic from users, operators should not just deploy more and more base stations to increase the capacity since this will increase the CAPEX and OPEX of operators. Moreover, it is necessary to avoid that the reduction of the cost of deployment and maintenance of base stations outweighs their revenues. Thus, a compromise must be made to achieve the tradeoff between the WiFi and the cellular networks. To study how much economics benefits can be generated due to delayed offloading, Lee *et al.* [71] modeled the interaction between a single provider and users based on two-stage sequential game. They quantify the benefits of delayed WiFi offloading in various aspects by conducting trace-driven numerical analysis. They use two traces, each of which tells us the information on cellular data usage and WiFi connectivity. This work shows that the increase ranges from 21% to 152% in the providers' revenue, and from 73% to 319% in the users' surplus.

Joe-Wong *et al.* [69] used empirical data to realistically study operators' saving from offloading. They propose a utility function and solve the revenue maximization problem, but we feel that more numerical investigations are needed. In addition to the operators' own WiFi APs, cellular operators may lease the third party wireless access point as a supplementary approach. To study the economic interaction between cellular network BSs and third-party WiFi or femtocell APs, Gao *et al.* considered a market-based solution, where macrocellular BSs pay APs for offloading traffic [70]. They use a non-cooperative game theory to figure out two factors: the amount of data being offloaded for each AP, and the payment given to each AP. In their proposed game, the BS proposes market prices, and accordingly the AP responds with the amount of data traffic it is willing to offload. They compare their outcome with two other classic market outcomes: outcomes of perfect competition market (no price participation), and outcomes of monopoly market (no price competition). They further analyze the impact of price participation and competition on the market outcome, shedding light to different aspects. Nevertheless, an interesting direction is to study the offloading market under incomplete information, such as what Iosifidis *et al.* [63] did. It has been proved to incur minimum communications overhead.

Thus, previous attempts provide that WiFi offloading can significantly benefit cellular operators [53]. However, in order to encourage users to offload their traffic to WiFi networks, cellular operators should give discounts or rebates to stimulate users to use the delay tolerance. For the benefit of operators, the incentive mechanism has to ensure that the total rebate to users is minimal. In fact, a novel incentive framework was proposed as a tradeoff between operators and users, which is called the Win-Coupon [64], [65]. It weighs the amount of traffic being offloaded and QoS. Note that the QoS varies in line with the different delay tolerance. The extent of the potential tolerant delay and the capability of offloading are taken into account, and both of them together determine the priority of the users to offload. The principle is that users with the ability to tolerate higher latency and greater potential of offloading should be given a higher priority.

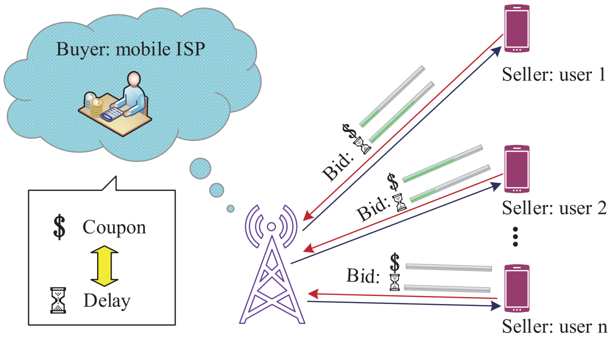


Fig. 4. Cost for Operators: The main idea of bids collection. Win-Coupon utilizes the bids collection module to infer users' delay tolerance [64].

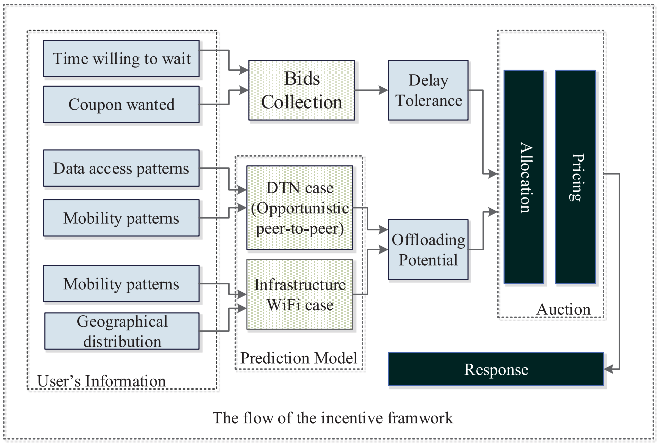


Fig. 5. Cost for Operators: Win-Coupon: the incentive framework for operators.

For the benefit of operator, to achieve the minimum incentive cost offered to users, the key is to take into account two factors: the delay tolerance and offloading potential of the users. First, in order to collect user's information regarding delay tolerance, an incentive framework based on reverse auctions [104] is adopted in the Win-Coupon. It proposes the reverse-auction scheme to stimulate users to offload their traffic. In other words, users auction their tolerant time of delay to cellular network operators in order to obtain the discount from operators, which is called coupon. Thus, as shown in Fig. 4, operators act as buyers and leverage coupons to pay for users' bids. Second, the evaluation on how much traffic the potential users can offload is based on two aspects. One is the size of the traffic required and the second is the probability of that a user passes by WiFi hotspots. Thus, to evaluate offloading potential of the users accurately, two different models for DTN and Infrastructure WiFi are considered.

As shown in Fig. 5, the main idea of Win-Coupon as an incentive framework is presented. To predict the delay tolerance, the users' bids information including the time that the users are willing to wait and coupon required by the users are collected to evaluate the tolerance. On the other hand, to predict the offloading potential, prediction models are classified into two cases including Infrastructure WiFi case and DTN case. Different users' information is collected to predict the offloading potential, respectively. In the allocation step, both the delay tolerance and the offloading potential are utilized to select the winners from the bidders. Then, in the pricing steps, operators determine the coupons they will offer to the winners. In the response step, the offered coupon and assigned delay will be sent back to bidders as a response.

In the Infrastructure WiFi case, users' mobility patterns and the geographical distribution of hotspots are used to predict the potential of offloading with Semi Markov models. In the DTN case (opportunistic peer-to-peer case), the random analysis based on access and mobile-based approaches are used to predict the potential of offloading. Then, as shown in the auction box in Fig. 5, the cellular network providers select the winners from the bidders to execute the allocation algorithm, along with the delay tolerance inferred by bids collection. Actually, the allocation problem is to determine the optimal solution which minimizes the total incentive cost, given an offloading target.

Although simulation results show that the nice efficiency and practical use of this framework within current prediction models, none of a single case can well support hybrid and complex scenarios. The hybrid framework in which the DTN case is well integrated with infrastructure WiFi case needs further research. In addition, a joint prediction model is needed to support the hybrid framework. Moreover, more advanced prefetching and caching schemes remain unclear.

2) Users' Perspective: In order to make it possible for users to make a tradeoff between cost, throughput quality and delay tradeoffs for different applications, Im *et al.* developed a system for cost-aware offloading: Adaptive bandwidth Management through User-Empowerment (AMUSE) mechanism [77]. This mechanism is a user-side WiFi offloading system. It is notable that users may not be willing to tell how long they can wait for traffic offloading for every application every time. The target of this mechanism is to provide users with an automated WiFi offloading decision which can intelligently offload data traffic on the basis of the preference settings made by users beforehand. This mechanism utilizes a utility maximization algorithm to take into account a users' throughput-delay tradeoff and a cellular budget constraint. A concrete measure of user's tradeoffs is used to mathematically formulate the user's offloading decision problem. The measure utilizes the utility function $U_j(p, t, r, s)$ to denote the utility of application $j \in J$ in period i, given that the day is divided into several discrete periods of time. Thus, the expression of the utility function takes into account four parameters: p (the per-volume price of the cellular network), t (the amount of time the session is deferred), r (the bandwidth speed at which the session is completed), s (the size of the session). So the incentive is to maximize the sum of the expected utility from WiFi and cellular networks.

The key of the bandwidth optimizer is to decide how long different applications should wait for WiFi in line with sessions usage prediction and WiFi access prediction as shown in Fig. 6. This mechanism is mainly realised by two main components: the bandwidth optimizer and the Transmission Control Protocol (TCP) rate controller. The bandwidth optimizer is devised to make offloading decisions for the user by enforcing

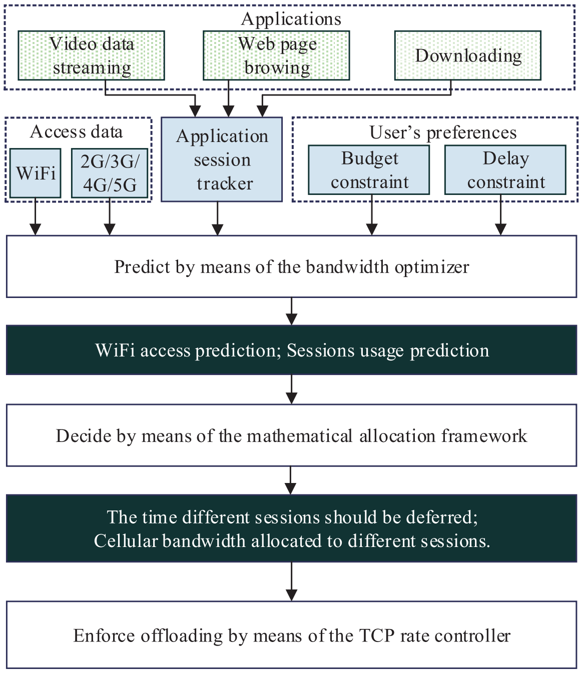

Applications
Video data streaming
Web page browing
Downloading
Access data
WiFi
2G/3G/4G/5G
Application session tracker
User's preferences
Budget constraint
Delay constraint
Predict by means of the bandwidth optimizer
WiFi access prediction; Sessions usage prediction
Decide by means of the mathematical allocation framework
The time different sessions should be deferred; Cellular bandwidth allocated to different sessions.
Enforce offloading by means of the TCP rate controller

Fig. 6. The main idea of AMUSE mechanism [77].

a utility maximization algorithm. The bandwidth optimizer consists of three components: the Future Usage Predictor, the WiFi Connectivity Predictor and the utility maximization algorithm. The Future Usage Predictor utilizes previous information to calculate the size $s_j(k)$ of each application type $j \in J's$ usage in each period k. The WiFi Connectivity Predictor utilizes a second-order Markov chain for the location prediction, which has been shown to be highly accurate [105]. Thus, the probabilities of WiFi accesses $p_l^{k+2}(l_k l_{k+1})$ for a user at location $l \in L_{k+2}$ during period $k+2$ can be predicted, given his locations l_k in period k and l_{k+1} in period $k+1$. Based on the results of the Future Usage Predictor and the WiFi Connectivity Predictor, the utility algorithm can calculate the expected utility of waiting for WiFi. The TCP rate controller on UEs is devised to enforce the cellular bandwidth allocations for those applications automatically along with the offloading decisions made by the bandwidth optimizer. To update the prediction modules, the application level session tracker is used to measure the actual usage for each application. Moreover, the measures of the actual usage for each application are displayed on the User Interface (UI) of UEs, thus users can set the rate of each application accordingly.

Researchers prototyped an AMUSE system on windows 7 tablet with this idea and evaluated the performance of the bandwidth optimizer in AMUSE. In the implementation, the Windows Filtering Platform (WFP) is used to track the application usage, and a user-side TCP rate controller algorithm is implemented. In this evaluation, the performance of AMUSE is compared to two main traditional offloading algorithms: on-the-spot offloading and delayed offloading. The amount of traffic offloaded is larger with AMUSE than it is with on-the-spot on average. This is because AMUSE allows delay-tolerant sessions to wait for WiFi accesses. On the other hand, the amount of spending is smaller with AMUSE than it is with on-the-spot offloading. The relationship between those three mechanisms can be well derived from Fig. 7. Moreover, in view of utility value calculated from the evaluation, AMUSE shows the best performance due to the utility maximization algorithm. So, it has a good performance in terms of the tradeoff between users' cost and the time the sessions are deferred. In general, unlike previous work related to delayed WiFi offloading, AMUSE provides the user-side TCP rate control algorithm. One of AMUSE's contributions is the automated mechanism, in contrast to some commercial applications that require users to manually specify the desired rate. However, it requires users to utilize computers to assist in managing offloading. What is more, it does not take into account the energy saving for UEs.

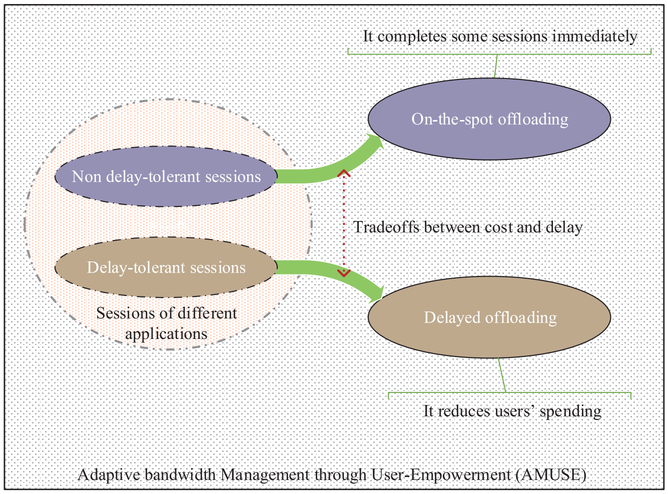

It completes some sessions immediately
On-the-spot offloading
Non delay-tolerant sessions
Tradeoffs between cost and delay
Delay-tolerant sessions
Sessions of different applications
Delayed offloading
It reduces users' spending
Adaptive bandwidth Management through User-Empowerment (AMUSE)

Fig. 7. AMUSE: it trades off between reducing users' spending by offloading delay-tolerant traffic and completing some non delay tolerant sessions immediately [72], [77].

Further work is still needed to investigate how users can obtain profits from WiFi offloading. Lee *et al.* [72] proposed a mobile tethering-based cooperative network which can make users obtain certain profits when users participate in this network. This cooperative network allows each user's data demand to be allocated to others, minimizing the total cellular traffic cost. Lee *et al.* model the cost problem, and maximize the total utility function of all participated users to determine the tethering price. Their simulation shows the reduction of total cellular traffic cost and finds the optimal tethering prices with different traffic demands. In addition, cache and prefetch mechanism are also interesting methods to minimize users' cost. El Chamie *et al.* [78] considered a scenario where users can download popular content from WiFi-enable caches deployed in an urban area. To figure out what contents should be put in each cache, El Chamie *et al.* make an attempt to find the optimal distribution of contents and propose two different cache coordination algorithms to achieve geographic fairness. They take into account content's popularity in different geographic places. Thus the average cost per requested content is minimized by an optimal replication and placement of content in caches. Nevertheless, investigating the effect of traffic dynamicity in their algorithms is still an interesting direction.

3) Summary: Previous studies did not show how much economic gain the provider and users can obtain, Lee *et al.* [71] quantified the economic gain of delayed offloading by fully considering interaction between mobile network operators and subscribers. This method's disadvantage in this study is that only the upper-bound of the benefits was given, and realistic scenarios between multiple providers and misinformation of user profiles are not solved yet. The Win-Coupon approach [64], [65] provided a novel incentive framework to motivate users to leverage their delay tolerance for cellular traffic offloading. This approach can effectively capture the dynamic characteristics of users' delay tolerance within reverse auction. Also, The Win-Coupon approach well guaranteed truthfulness, individual rationality, and low computational complexity. However, the realistic scenario is a hybrid case and Win-Coupon's framework operating in this hybrid case is unclear.

The AMUSE approach [77] well navigated the tradeoff between cost, throughput and delay when users' offloading decisions are made. However, this approach did not well address instantaneous variation of network conditions and also did not support cooperative network and prefetch mechanism. Lee *et al.* [72] well considered the poor coverage of current WiFi networks by a tethering-based cooperative approach, but the impact of computation complexity on implementation is unclear. El Chamie *et al.* [78] presented the first study to explore the problem of unfairness that might result from content placement for prefetch mechanism. However, the effect of traffic dynamicity and the way of traffic dynamicity interacting with the developed algorithms are not presented. Prefetch mechanism can significantly reduce users' cost, but it requires operators' support. Overall, future directions regarding cost are to enhance all the existing mechanisms within seamless experience and concurrent connections.

B. Energy

With the popularity of the idea of energy saving, green communication has attracted more attention in the last few years. This is also important for WiFi offloading schemes. It is necessary to investigate how energy consumption can be saved in heterogeneous network for operators. On the other hand, users may be willing to accept WiFi offloading when they can increase the service time by saving the energy consumed on mobile devices. In order to decrease the energy consumption from operators' and users' perspectives, many analyses and experiments have been conducted to study the energy saving for WiFi offloading in various scenarios.

1) Operators' Perspective: To offload data traffic from traditional cellular networks, operators usually deploy more small cell BSs consisting of picocell BSs, femto-BS, and WiFi APs. However, deploying small cell BSs requires backhaul networks which connect the small cell BSs and the mobile core networks. This may neutralize the energy efficiency. Thus, cellular operators can lease the WiFi APs deployed by the third party such as Internet Service Providers (ISP). Paris *et al.* [79] proposed a new and open market to allow cellular operators lease the third party WiFi access network. To select the cheapest APs and offloading the maximum amount of traffic, they formulate the offloading problem as a reverse auction between each AP owner and cellular operator. They showed that a lot of energy consumption can be reduced by the opportunistic utilization of third party WiFi APs. The energy and spectrum efficiency can be enhanced because APs deployed by ISP are usually closer to users. Han and Ansari [80] proposed a novel mobile traffic offloading scheme referred to as energy spectrum trading (EST). They investigate extra bandwidth allocations to WiFi APs and users-BSs associations to minimize the energy consumption of BSs. In fact, it is difficult to minimize the energy consumption of the macro BSs to achieve optimal traffic offloading. They make an attempt and propose a heuristic power consumption minimization (HPCM) algorithm to approximate the optimal solution with low computation complexity. Their simulation showed that at least half of the energy consumption can be saved.

Cellular operators can also leverage already existing WiFi mesh networks to establish cellular-to-mesh (C2M) data offloading. The mesh networks are built and managed collaboratively by users. Cellular can lease these networks to reduce investment and energy consumptions. Considering the energy consumption on the cellular network BSs, Apostolaras *et al.* [81] proposed an analytical framework to determine which users should be offloaded. They simulate the operation of the LTE-A network and conduct testbed experiments for the mesh network, proving that significant energy savings can be achieved.

Note that, other than being offloaded to the WiFi network, the traffic can be offloaded from the cellular network to femtocells that are purchased and deployed by users, which can relieve the load of the cellular network without any extra cost to operators and can also save energy consumption on base stations. On the other hand, the cellular systems may have a large path loss and unnecessary energy consumption when some UEs are indoor. Thus, it is necessary for operators to utilize emerging access networks in an efficient and collaborative manner to determine which wireless access network the traffic should be offloaded to. The integrated femtocell-WiFi network [106] is a choice for integrating multiple advantages of two networks. The scheme can migrate traffic from the cellular network to the WiFi network and reduce the load so as to avoid congestion. On the other hand, this algorithm can reduce energy expenditure for operators. The target is to utilize WiFi and femtocell to reduce energy consumption of the cellular networks. This can be done by switching down the unused base stations intelligently. However, it is crucial to deploy those radio accesses successfully under a mixed spectrum of unlicensed and licensed bands. AlQahtani proposed a new cognitive-based connection-level admission control scheme [107], which is done with access retrial for a femtocell network. AlQahtani utilized the retrial phenomenon policy to reduce femtocell users' loss probability, where the performance of loss probability and throughput compared to traditional femtocell operations had been significantly improved.

In addition, the interference of femtocells may be managed by cellular operators, but the interference management between WiFi accesses is still a challenge. The interference has been extensively investigated to integrate femtocell with WiFi network seamlessly. When the interferer and the victim belong to different network layers, the interference is called cross-tier interference. Yeh *et al.* [108] showed that interference management schemes are critical for reducing the resulting cross-tier interference. Moreover, it is significant to allocate traffic efficiently between those access networks. The traffic management method for saving energy has been discussed from the operators' perspective by some researchers [109]. They argued that significant energy was saved through power saving modes. But the seamless handover and protocols translation were still unclear. What is more, the information exchange interfaces are managed manually. This does not help reduce the expenditure for operators in practice.

2) Users' Perspective: In general, when the distance between the WiFi AP and the UE increases, the WiFi signal strength decreases while the energy consumed increases. This trend also applies to cellular networks. On the other hand, the energy consumption of both WiFi and cellular networks increases with the network load, but the coefficients of variations for the two signals are different. Thus, the function curves will intersect when the load reaches an appropriate value, and this value is adopted as the threshold of decision mechanism that determines the time when to hand over and which network should be selected.

Considering these characteristics, researchers try to describe the energy consumption in a generic model which is a function of load and bit rate. The energy consumption at different bit rates or load are widely measured, and a function of the curve is obtained in line with the experimental results which turns out to be nearly linear, and the results can be used to determine parameters to infer the mathematical model in both cases. Then, the two models can draw a comprehensive unified mathematical model to describe the energy [83]. The energy consumption is given by

$$E = \frac{C}{R + D} + (M - 1000) \cdot A \tag{1}$$

where E represents the energy consumed in Joules, R denotes the bit rate in Kbps and M represents the amount of load, while A, C, D are constant parameters that vary with WiFi and cellular networks.

It is attractive to use two wireless interfaces simultaneously for traffic offloading. With the release-10 of 3GPP [110], IP Flow Mobility (IFOM) is one promising mechanism to enable a UE to maintain two concurrent connections with cellular BSs and WiFi APs [111]. Thus, it allows selected IP flows to be offloaded via multiple interfaces simultaneously, dividing it into multiple sub-flows. The key advantage is that it supports seamless shifting: an IP flow can be shifted to different interfaces without disrupting an ongoing communication. The challenge is to reduce energy consumption for concurrent connections. The majority of the state-of-the-art focus on downlink offloading. To investigate the energy saving for uplink offloading using IFOM, Miliotis *et al.* [91], [97] made an attempt to propose two uplink offloading algorithms for IFOM. In their first algorithm, UEs with heavy traffic are given priority in accessing the WiFi APs. In their second algorithm, a proportionally fair bandwidth allocation over the data volume demands from UEs is presented. In their simulation, they present the limitations of IEEE 802.11 DCF (Distributed Coordination Function) uplink access [112] by comparing it with their two uplink offloading algorithms. For IFOM uplink offloading, they show that it is significant to improve the uplink access scheme of WiFi in terms of saving energy consumption on UEs.

It is still not clear how to provide IP mobility in cellular networks with current protocols such as Proxy Mobile IPv6. To support network-based mobility management, Proxy Mobile IPv6 [113] was proposed by the IETF (Internet Engineering Task Force). It extends Mobile IPv6 with additional functional entities: the mobile access gateway, and the local mobility anchor. The gateway is responsible for tracking the mobile node in the radio access links. The anchor is responsible for collecting routing information for mobile nodes. To reduce the energy consumption, Sanchez *et al.* [85] presented some network-based IP flow mobility extensions, and use experimental measurements to investigate how energy savings can be achieved on users' terminals. Their experiments provide that WiFi shows better performance of energy efficiency and throughput, compared to cellular interface.

It is further interesting to combine multiple concurrent data streams with device-to-device techniques such as WiFi direct and Bluetooth. Sharafeddine *et al.* [86] considered a cooperative network scenario where a smartphone uses its WiFi and Bluetooth interfaces simultaneously to offload traffic via local D2D links. They design and implement a testbed for mobile cooperative multimedia distribution to evaluate performance in terms of energy consumption. Their evaluation shows that Bluetooth is more energy efficient that WiFi-direct on D2D links. In addition, WiFi is more energy efficient than 3G on the long range link for the considered experiments. This result presents typical values that can be encountered in real scenarios. However, using multiple interfaces incurs the largest value of energy consumption. In fact, a smart allocation mechanism is needed to allocate different traffic to different interfaces intelligent, considering instantaneous conditions and historical information of different interfaces. Furthermore, just as what Siris and Anagnostopoulou [84] did, it would be an interesting approach to investigate energy savings by exploiting mobility prediction and prefetching to enhance WiFi offloading schemes.

3) Summary: Approaches [79], [80] leasing the WiFi APs deployed by a third party are of high feasibility, but it is difficult to minimize the energy consumption of BSs to achieve optimal traffic offloading. By utilizing WiFi mesh networks, [81] gave a full consideration to energy saving and profit sharing for users' participation, but the final operator's revenue from users is not clear. By improving the femtocell-WiFi network, AlQahtani [107] successfully enhanced the performance of loss probability and throughput, but further studies are needed to investigate the influence of interference and traffic management on the energy consumption.

By measuring the impact of rate/load size on the energy consumption, the experimental approach [83] simplified the operations on UEs regardless of network conditions, user preferences, etc. The disadvantage of [83] is that it required manual switching and the accuracy needs to be enhanced by a non-linear approximation. Considering the trend of concurrent connections in the future, Miliotis *et al.* [91], [97] well extended the state-of-art to uplink scenario with a fair bandwidth allocation algorithm. The disadvantage is that the impact of computation complexity on realistic UEs is unclear. Sanchez *et al.* [85] defined some extensions of Proxy Mobile IPv6 that benefit both the network operator and the end user in terms of the energy saving. However, energy efficiency of real devices in different operating systems and architectures is not clear yet. Siris *et al.* [84] exploited mobility prediction and prefetching and well evaluated the energy consumption and throughput, but a prototype has not been done. In conclusion, a future direction for operators is a systemic approach to trading off between throughput and energy saving, while the future direction for users is to implement a smart and automatic allocation mechanism on UEs within concurrent connections.

C. Rate

In this section, we utilize the concept of rate to describe the performance in terms of average completion time of demanded data transmissions in WiFi offloading, which can determine the QoS directly. Actually, several factors jointly contribute to practical rate such as information transmission rate of a radio access network, switching delay, practical load condition, and congestion. Considering the incentive from which perspective, the state-of-the-art can be classified into two categories as discussed below.

1) Operators' Perspective: To improve the information transmission rate, deploying advanced radio access techniques is a key approach for operators. While for WiFi offloading which mainly exploits existing radio access techniques, the key that operators need to consider is to avoid congestion and keep a balance between different radio access networks' loads. Previous proposed schemes focus on alleviating the cellular congestion by offloading the data traffic to WiFi as much as possible, but without systematic considerations of the network congestion, load balancing, and completion time. The WiFi offloading may not perform well when congestions are not considered in WiFi offloading.

Cheung *et al.* [6] formulated the congestion-aware selection problem as a network selection game (NSG) in an integrated cellular-WiFi system. NSG incorporates some practical considerations that include user mobility, WiFi availability, switching time, and cost of switching networks based on usage. Their simulation shows that this scheme could significantly balance the load of the overall networks, improving the throughput and reducing the handover delay based on WiFi-3G cellular hybrid systems. However, the scenario, where WiFi networks is integrated with LTE networks that have a different performance of delay, remains unclear.

Note that small cell base station (SCBSs) are increasingly deployed within a macro sector to make cell sizes smaller. SCBSs are becoming capable of leveraging cellular network RATs and WLAN RATs simultaneously, operating on licensed bands and unlicensed bands, respectively. An optimizing issue for SCBSs is how to smartly steer traffic from cellular networks to WLAN to optimize the performance and QoS, integrating the two RATs efficiently. Therefore, a smart offloading scheme should be capable of considering instantaneous network conditions and user's requirements and allow small cells and WLAN to learn their own optimal transmission strategies.

Inspired by the reinforcement learning theory [114], Simsek *et al.* [98] proposed a distributed traffic steering framework for small cells. This framework is engaged in self-organization process based on a cross-system learning framework. In a *cross-system learning framework*, SCSBs carry out the cross-system learning procedure on licensed and unlicensed bands jointly to learn their long-term metrics and derive their optimal transmission strategies in order to balance loads between cellular networks and WLAN to avoid congestions. The cross-system learning framework allows SCBSs to select suitable subbands in licensed and unlicensed bands and allocates the selected subbands to one UE after a successful attempt of accessing the unlicensed bands so as to balance the loads between the two network modes. It is provided that the cross-system learning framework can improve the average throughput of UEs by steering traffic intelligently and dynamically on licensed bands and unlicensed bands.

Instead of serving an arbitrary number of UEs, the cross-system learning framework can be coupled with a *traffic-aware scheduler*. SCBSs carry out the traffic-aware scheduling procedure so as to schedule their UEs by allocating different priorities to UEs according to UEs' information such as transmission conditions and files information. Thus, SCSBs perform resource block allocation to allocate their selected subbands to the scheduled UEs. To reduce the completion time of transmission, Hu *et al.* [99] proposed traffic aided opportunistic scheduling (TAOS) by considering both the file size information and channel variation to schedule UEs. Based on the works related to TAOS, the *traffic-aware scheduler* makes schedules by considering instantaneous channel conditions, congestion levels, completion time, file sizes and UEs' service types. Integrating "picking up the users with the shortest time" [115] with "picking up the users on the channel peak" [116], [117], SCSBs pick up their UEs to reduce the total completion time. Thus, UEs with short completion time traffic are steered to licensed bands while UEs with delay tolerant traffic are steered to unlicensed bands. Simulation results demonstrate this point is true by steering traffic intelligently and dynamically on both licensed and unlicensed bands. Furthermore, significant improvement can be achieved when this cross-system learning framework is coupled with a traffic-aware scheduler. However, this framework needs further investigation on users with high mobility, and this framework does not give more details regarding interference management.

Furthermore, it is interesting to study a load coupled network for WiFi offloading analytically. Siomina and Yuan [118] presented a theoretical analysis of cell load coupling for LTE networks. They employ the load coupling equation derived from a SINR model that takes into account the load of each cell. For a given demand, the load of each cell depends on a non-linear manner on the load of other cells. This loading coupling model shows a good approximation for more complicated load models. However, their work just focused on the demand to be served in cellular networks. To extend this idea to WiFi offloading and achieve load coupling in a heterogeneous network, Ho *et al.* [100], [101] formulated a utility-maximization problem. They optimize the demand to be served in a heterogeneous network to maximize the utility function. Considering practical implementation, they also propose a strategy to constraint the load to some maximum value. Further work is still needed to consider more factors such as energy saving and user-network association.

In particular, to achieve load balancing in mobile adhoc networks (MADNETs), Kumar and Ramachandram [119] investigated load balancing for MADNETs. For a MADNET, they demonstrate that Genetic Zone Routing Protocol (GZRP) is very useful in treating congestion in cellular-WiFi networks. To investigate load balancing routing for 3G-WiFi offloading networks, Budiyanto *et al.* [120] combined GZRP with VHO, showing the best performance. It not only improves the throughput but also reduces the handover delay. VHO can well process the handover from the cellular network to the WiFi access, but it still needs to be combined with GZRP as the routing algorithm in order to avoid congestion when considering WiFi APs. Nevertheless, further work is needed to consider how it affects the LTE-WiFi offloading network.

2) Users' Perspective: From user's perspective, the direction is to study how the average rate per user can be improved. To investigate the average rate for users analytically, Singh *et al.* [102] made an attempt to figure out the optimum fraction of traffic being offloaded. They present a tractable model to analyze the effect of offloading. This model considers a heterogeneous network that consists of M different RATs, each deploying up to K different tiers of WiFi APs. They define two concepts: SINR coverage, and Rate coverage. SINR coverage is defined as the probability that a randomly located user has SINR greater that an arbitrary threshold. Rate coverage is defined as the probability that a randomly located user has greater rate than an arbitrary threshold. Rate coverage takes into account SINR, load, RAT's condition, and users' requirements. They formulate the offloading problem into two incentive functions. It is proven that there exists an optimum percentage of the traffic that should be offloaded for maximizing the rate coverage. Nevertheless, it is interesting to extend this work by investigating the coupling of AP queues. In addition, further work is needed to extend the inter-RAT offloading to inter-tier offloading in one RAT. Furthermore, it is very interesting to analyze how a simultaneous transmission scheme affects average rate per user. Galinina *et al.* [103] proposed an architecture to support dynamic data flow splitting across integrated dual-RAT infrastructure in ultra-dense small cells. They propose a novel analytical methodology for ultra-dense LTE/WiFi heterogeneous networks and deliver a comprehensive analytical model. This unprecedented analytical framework has significantly advanced the state-of-the-art in the offloading field. In fact, this work outlined a truly integrated LTE/WiFi HetNets for the design of emerging 5G systems.

3) Summary: Previous literature did not consider the network congestion, switching penalty, and network pricing in data offloading. Cheung's work [6] is the first study to focus on the congestion-aware network selection and data offloading problems with multiple heterogeneous users. However, this approach is restricted to the scenario where the cellular network is available to all the UEs at all possible locations all the time. A more general scenario with random mobility patterns is not addressed. The traffic steering framework employing a traffic-aware scheduler in [98] can take full advantage of both WiFi and cellular network to support seamless experience. The advantage of cross-system learning framework is alleviating of instantaneous inputs. The disadvantage of this approach is that it only was applied to low-mobility UEs. Moreover, interference management and interplay between mobility and cell association are ignored. Unlike the framework [98] where SCBS allocates suitable subbands to UEs, SCBS in the approach of TAOS [99] steers UEs to licensed/unlicensed bands according to UEs' completion time. The advantage of TAOS is that it can reduce the total completion time, and perform well in heavy-loaded wireless networks. However, it leads to extra traffic transmitting due to file size information and channel state information. Reducing completion time may damage performance of the total throughput. As for the theoretical analysis of load-coupled networks, Ho *et al.* [100], [101] extended the idea in [118] from cellular networks to heterogeneous networks and made it meet the demands of practical implementation. However, energy minimization and user-network association is ignored.

As for the theoretical analysis of the average rate for UEs, Singh *et al.* [102] well considered SINR, load, RAT's condition, and users' requirements under a flexible association model. Singh's work was the first to study rate coverage in the context of inter-RAT offload. However, the coupling of AP queues and the inter-tier offloading within a RAT are ignored. Also, a non-linear approximation is needed to improve the area approximation for the association regions. Singh's approach did not solve the problem of data loss during switching and simultaneous usage of multiple RATs. Galinina's approach [103] enabled dynamic data flow splitting across integrated dual-RAT infrastructure and well considered the load of each small cell varying significantly over time and space. Galinina's approach significantly shortened end-to-end delays and lowered small cells deployment costs. Unfortunately, in this work, the existing LTE/WiFi architecture needs to be updated and a new entity alien access gateway should be deployed on the interface between WiFi and EPC. Furthermore, energy savings are ignored in this work. In conclusion, the future direction of improving rate is to utilize instantaneous information [6], [99], [100], [102], [103] or learning procedures [98] to support seamless experience with simultaneous usage of RATs at low cost of energy consumption.

D. Classification of Related Individual Approaches

In addition to the basic incentive of capacity, three particular incentives (cost, energy and rate) have been discussed

TABLE X
CLASSIFICATION OF RELATED INDIVIDUAL APPROACHES

Incent.	Persp.	Issues	Approaches	Advantages	Disadvantages
Cost	Operators	Modelling benefits of delay	Jo. Lee *et al.*[71]	Benefit for operator and users	Misinformation of user profiles
		A incentive framework	Zhuo *et al.*[64]	Low computational complexity	Lack of realistic scenario
	Users	Adaptive bandwidth	Im *et al.*[77]	Navigate the tradeoff	Without cooperation/prefetch
		Cooperative approach	Ji. Lee *et al.*[72]	Independent of APs' coverage	Lack of implementation on UEs
		Cache/Prefetch mechanism	Chamie *et al.*[78]	Solve the problem of unfairness	Require operators' support
Energy	Operators	Lease the WiFi APs	Han *et al.*[80]	Spectrum efficiency	Without considerating throughput
		WiFi mesh networks	Apos. *et al.*[81]	Users' participation	Unclear operators' final revenue
		Femtocell-WiFi network	AlQa. *et al.*[107]	Loss probability and throughput	No Interference/traffic management
	Users	Impact of rate/load size	Taleb *et al.*[83]	Simplify the operations on UEs	With simple manual switching
		Energy saving for IFOM	Miliotis *et al.*[91]	Uplink scenario	Unclear result of realistic UEs
		Experimental measurements	Sanchez *et al.*[85]	Extend Proxy Mobile IPv6	Restricted platform/architectures
Rate	Operators	Congestion-aware NSS	Cheung *et al.*[6]	Multiple heterogeneous users	Restricted scenarios of networks
		Steering of subbands	Simsek *et al.*[98]	Alleviate instantaneous inputs	Only for low-mobility UEs
		Steering UEs to WiFi/LTE	Hu *et al.*[99]	Heavy-loaded wireless networks	Extra traffic transmission
		Analysis of load coupling	Ho *et al.*[100]	Practical implementation	Without energy saving/association
	Users	Theoretical analysis	Singh *et al.*[102]	Inter-RAT offloading	Lack of inter-tier, Data loss
		Dynamic data flow splitting	Galinina *et al.*[103]	Shorten delay/lower cost	Not for current infrastructure

TABLE XI
TECHNIQUES ACCORDING TO FURTHER INCENTIVE: CONTINUITY

Further Incentive: Continuity			
Enhanced NSS	Relaying Mechanism	Multipath Mechanism	Collaborative Mechanism
Real-time Switch	D2D Communications	Multipath TCP	Opportunistic Offloading
[125]	[129][130][131][132][133][134][135][136][137][138]	[150][151] [152] [153][154][155]	[158][159][160]
Context Aware Handoff	P2P offloading	IFOM (IP Flow Mobility)	MADNET
[126][127]	[140][141][139][142]	[85][91][95][97]	[157][89]

in the previous three subsections, respectively. Eventually, to better compare related individual approaches from different incentives, advantages and disadvantages of related individual approaches are listed in Table X.

V. FURTHER INCENTIVE: CONTINUITY

It is apparent that the WiFi offloading scheme may make UEs switch radio access network frequently. This may incur disruptions to ongoing communications, damaging users' satisfaction significantly. What is worse, a high dynamic of mobile communication environment and a fluctuating wireless channel will incur frequent disruptions to ongoing communications, especially for vehicular scenarios. Thus, it is very important to investigate how to alleviate disruptions and maintain continuous communications. The state-of-the-art can be classified into four categories. The first is to enhance the NSS to reduce handoffs and interruptions. The second is to use relay mechanism between ad-hoc networks such as device-to-device communications. The third is to split data flow into sub flows and transmit them simultaneously via multiple paths. The fourth is to provide collaborative mechanism between operators and users. A detailed classification is shown in Table XI.

A. Enhanced NSS

1) Real-Time Switching Between Interfaces: Contemporary devices such as Android 4.0 phones are capable of accessing both cellular as well as WiFi. They are also allowed to switch interfaces by simply turning on an interface while turning off the other interface manually, which does not support seamless switching. Simply switching between multiple interfaces is not able to be applied into ongoing TCP sessions since switching between network interfaces is bound to incur interruption and data loss. In order to address the continuity issue, several research groups [121]–[124] attempted to deal with interruptions during switching between wireless interfaces. However, their approaches either need to extend the current infrastructures or require novel protocols instead of existing protocols. Unlike delay tolerance offloading, real-time switching approaches explore dynamical switching to deal with the interruptions between interfaces. In order to optimize the performance such as battery time, data offloaded and throughput, MultiNet [125] was proposed to utilize three switching policies to switch interfaces in real-time. MultiNet is fully client-based without requiring additional infrastructure support or any changes to existing protocols, which is of high feasibility. Taking the characteristics of TCP sessions and UDP (User Datagram Protocol) sessions into consideration, MutiNet deals with TCP sessions and UDP sessions separately to avoid interruptions. In a real implementation such as Android platforms, the switching module is achieved by Application Program Interface (API) with a *C*++ method to switch the connection to a new interface.

MultiNet architecture is composed of three modules including switching engine, monitoring engine and selection policy.

The switching engine is responsible to perform the switching between cellular networks and WiFi access networks. The monitoring engine is responsible for monitoring and storing system variables such as the amount of data traffic, connectivity status, the status of ongoing TCP and UDP sessions, and the parameters of battery on mobile devices. These variables are necessary for switching decisions. In terms of the performance needed to be optimized, selection policies can be divided into three categories: energy saving, offloading ratio and performance. The aim of the energy saving policy is to minimize the energy consumption on mobile devices. In MultiNet, mobile devices will connect to the cellular network when it is idle. MultiNet utilizes dynamical switching to allow a mobile device to switch to the WiFi access network to offload data traffic once it detects that the amount of data traffic exceeds the predefined amount threshold. Further, a mobile device will switch back to the cellular network as soon as the time of connecting to WiFi exceeds a predefined time threshold. That is because the power of WiFi is prone to be significantly higher than that of the cellular network, while the energy consumption of WiFi is prone to be lower than that of cellular networks when those two type of access networks transfer the same amount of data traffic that exceeds a threshold. The aim of the offloading policy is to offload the data traffic as far as possible unless the RSS of the WiFi does not exceed the threshold, while the energy consumption is not efficient when the WiFi is idle. The aim of performance policy is to improve the throughput by switching to the interface with the highest bandwidth.

The TCP connections are bound to be interrupted when the interfaces are being switched, while most sessions utilize TCP instead of UDP. MultiNet deals with these two types of sessions separately. For UDP sessions, MultiNet does not wait for the ongoing session over old interfaces to be completed, which is bound to incur some packet loss of the in-bound traffic. For TCP sessions, MultiNet adds new routing table entries. When it turns on the new interface, it will not turn off the old interfaces occupied by TCP sessions until the mobile devices complete the ongoing TCP sessions over the old interfaces, ensuring that the ongoing TCP sessions will not be interrupted during switching. Thus, the difference between real time switching and multihoming is that multiple interfaces are turned on only when UEs determine switch from current interface to another one in real time switching. Once this transition state of real time switching is completed, only one interface is turned on. Some researchers [125] utilized Android-based mobile devices to implement three switching policies of MultiNet in a real-world scenario, which implemented the Android system by adding some non-existing modules without any changes to the existing network protocols. To control switching in Android, two useful methods can be utilized in the Android applications within *C++*, including *switchInterface()* and *useInterface()*. They demonstrated that the MultiNet system outperformed the state-of-the-art Android system either by saving energy consumption up to 33.75%, or nearly achieving the optimal offloading amount and throughput that can be achieved by "switching to WiFi whenever it is available" approach. In addition, MultiNet had successfully avoided the disconnections that occur in "switching to WiFi whenever it is available". Moreover, the energy saving mode is preferred in terms of the tradeoff between energy consumption and offloading amount since the energy consumption of *the offload mode* is substantially larger than that of *the energy saving mode* while the offloading amount of *the energy saving mode* is slightly smaller than that of *the offload mode*. In the offload mode, although WiFi networks are utilized to offload mobile data traffic, mobile devices will be connected to cellular networks to save energy consumption when mobile devices are idle.

2) Context Aware Handoff Mechanism: Since coexistence of multiple network interfaces is the trend of mobile communication systems, a user can select multiple networks and switch to the interface with the best performance. However, the experience may not satisfy users with poor network performance during the handoff process between multiple networks. The key problem is service interruption of applications.

In order to optimize the performance of services, the handoff schemes need to take more information into consideration such as the status of network access and kinds of applications running on the smart-phones, for instance. This is due to the fact that interruptions vary with different applications since the latter adopts different communication protocols or cache schemes. Generally, there are three main challenges for determining which application should switch its connection. The first is that the number of applications is enormous. Second, the usage of certain applications on smart-phones is unpredictable. Third, there will certainly be some new future applications that are currently not known.

In order to address these challenges, Context Aware Handoff Policy utilizing crowd-sourced application was proposed by Li *et al.* [126]. In addition, Howe [127] devised a way to allow subscribers to request contributions from a large crowd of people. To overcome the usage boundedness in this approach, the Context Aware Handoff Policy utilizes similarity features to divide the users into several crowds. Thus, users in the same crowd can share application data. This can be done by the context model based on Bayes classifier [128]. The main idea is shown in Fig. 8.

Context Aware Handoff Policy is composed of three parts. The first is monitoring a daemon which collects four kinds of context: network conditions, power data, traffic volume and application information. The second is the network selection scheme which is drawn from three metrics that include network performance, energy consumption, and cost. These metrics can be derived from the first part. This selection scheme is of high feasibility for its simple algorithm. The third one is the application context model based on Bayes classifier in order to know the probability of what type of network should be selected. In the function of Bayes classifier, four kinds of context are combined into a vector as the input parameter. In order to address the challenges of the uncertain usage and enormous kinds of applications, researchers proposed to utilize the crowd-sourced application data to construct the context model. This context model includes two similarity metrics: smart-phone usage similarity and network conditions similarity. Based on the computation of two similarity values by hierarchical clustering techniques, each

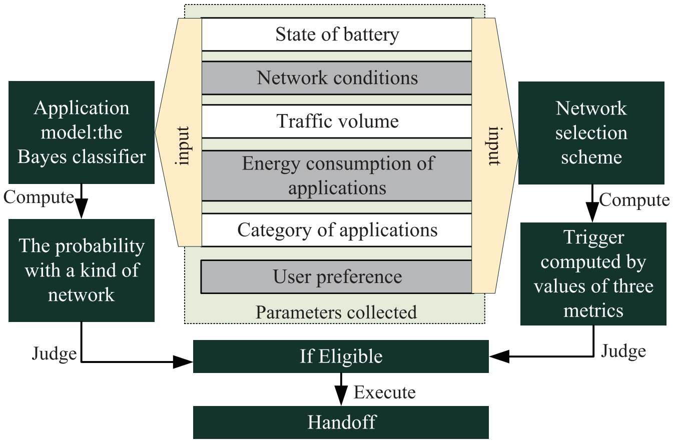


Fig. 8. The main idea of context aware offloading. After some parameters are collected, the Bayes classifier works out the probability with a kind of network and the Network selection scheme works out the trigger, respectively. Then the results are used to judge whether the condition is eligible for handoff or not.

crowd can be discriminated. In the evaluation performed by Li *et al.* [126], this proposed scheme achieves 25% energy consumption, 40% data offloading and 200% throughput. The reduction in the frequency of handoffs was achieved by two third compared to traditional approaches.

B. Relaying Mechanism

1) Device-to-Device Communications: It is common that there are many users around the WiFi infrastructures, which may conceal the users at a distance. What is worse, a WiFi network cannot track receivers even if they have transmitted data at one time via a WiFi network. Thus, the infrastructure WiFi path is weak considering the limit of transmission distance. On the other hand, in order to improve the capacity of a cellular network, it is promising for Peer-to-Peer technology to be integrated into LTE architecture. WiFi D2D offloading based on IEEE 802.11 protocol stack can effectively utilize short links between peers to deliver traffic directly instead of using an infrastructure path. This can improve the capacity and data rate significantly. It is notable that WiFi D2D offloading is based on the considerable density of WiFi devices and outperforms the infrastructure significantly when the interference is not severe.

However, a management scheme to avoid overload for WiFi D2D should be developed to steer the offloading volume. This can maintain the merit of performance such as energy efficiency and low delays. Evaluations showed that the throughput increases significantly when the offloading percentage of total traffic grows. In addition, the performance of D2D links decreases since increasing offloading traffic leads to increasing D2D links. This in turn contributes to some noise increase on WiFi bands. Moreover, the energy efficiency of D2D decreases as the offloading percentage grows. In addition, increasing offloading can significantly contribute to MAC delay denoting the time the packet takes to get through the MAC layer before being acknowledged. Therefore, the benefits of offloading cannot be guaranteed when non-controlled connections are established.

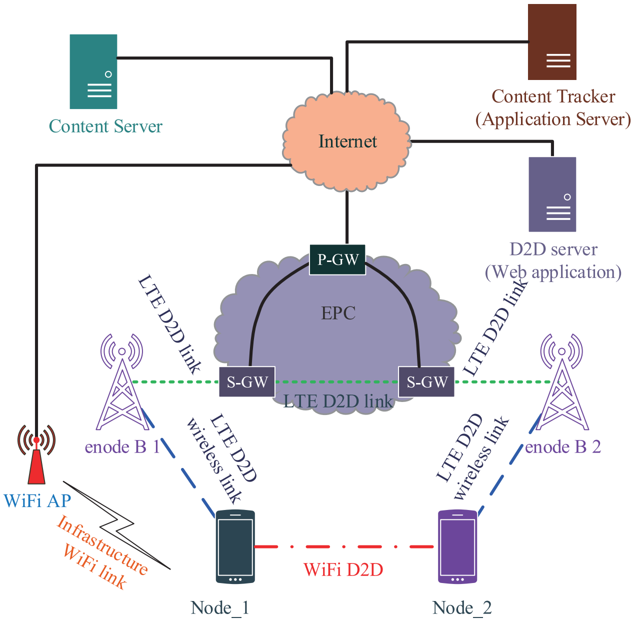


Fig. 9. WiFi D2D link compared to LTE D2D link.

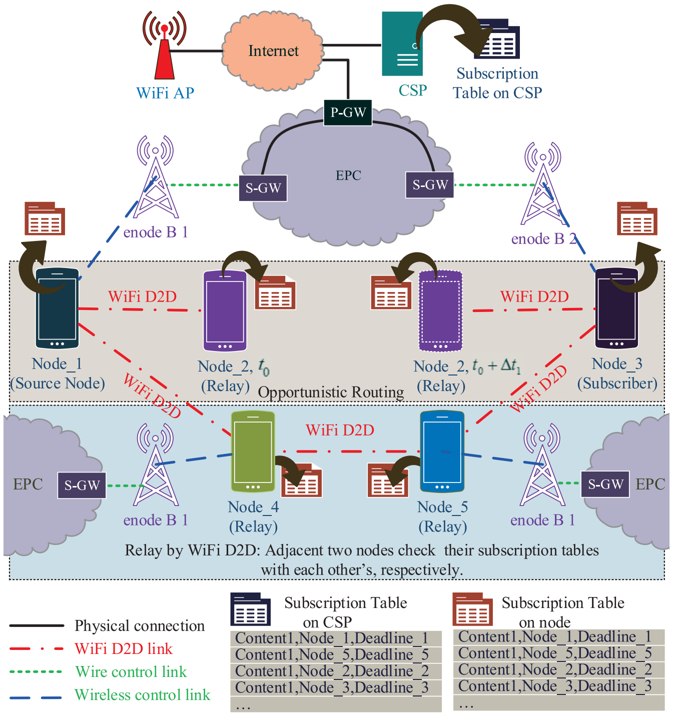
Recently, D2D communications integrated into the LTE network is devised to circumvent this problem. The LTE network can utilize its advantages to help manage connections between devices in this approach. According to the function of LTE networks, D2D communications are divided into two types: network-controlled D2D and network-assisted D2D. In network-controlled D2D, the LTE network is responsible for device discovery, D2D connection establishment and radio resource allocation for D2D connections. In network-assisted D2D, the LTE network is only responsible for device discovery and D2D connection establishment. Direct radio links between devices include LTE direct and WiFi direct in this integrated network. Consequently, there are two different kinds of D2D communications including licensed band D2D communication and unlicensed band D2D communication. As a licensed communication, LTE direct acts as an underlay to the LTE network. However, the pace of LTE direct is slow, while WiFi direct has been standardized and WLAN D2D protocol has attracted many researchers' attention in the last few years. Based on network-assisted and WLAN D2D protocols, a cellular network-assisted WiFi direct communication is a promising approach as discussed in recent literature.

In this framework, network assistance helps devices to discover other devices in proximity and establish WiFi D2D connections. As shown in Fig. 9, in addition to users' devices in this architecture, two main participants are a content tracker and a D2D server. A content tracker such as the application server is responsible for storing all available contents' locations from all registered devices and servers for requesting clients. The D2D server is typically deployed in EPC network, and is responsible for establishing clients' D2D connections in the transport layer. Based on the prototype of a web application composed of a web server and a web client [129], the evaluation of this prototype shows that WiFi Direct can significantly improve capacity and throughput of networks and decrease delay and energy consumption for clients compared to LTE direct.

There are still several challenges in the WiFi Direct approach. First, WiFi Direct needs a scheme to detect WiFi peers. Second, an algorithm is also needed to help determine whether and when the portion traffic should be offloaded through D2D links to avoid competition [130]. Third, a scheme needs to be developed to cooperate with the cellular network and determine which part of traffic will be transferred via WiFi D2D in order to ensure continuity of services [131]. Fourth, since data rates of different Direct Links in cellular controlled short-range communication system [132] can vary greatly, a scheme is necessary to determine the appropriate D2D WiFi link to achieve optimal user performance. Last, an appropriate power control mechanism can also help WiFi D2D improve energy efficiency [133]. All the above discussed schemes for WiFi direct can be categorized under network-assisted D2D offloading. Evaluation in [134] showed that D2D offloading improves both capacity and energy efficiency significantly even with the help of the simplest scheme, implying a promising way for improving performance of offloading in the future.

2) Peer to Peer Offloading: Other than direct transfers on smart-phones through Internet servers, as an opportunistic offloading scheme, Peer to Peer (P2P) WiFi communications is showing a significant importance recently. This can take the role of offloading schemes, which is based on the idea of WiFi Direct communications discussed earlier. So in addition to WiFi direct connections between two devices, more peers that can relay contents are added into the WiFi Direct communication architecture. This way, P2P WiFi can be considered as an extension of WiFi Direct. However, the drawback of P2P WiFi is that the data transfers between peers are not guaranteed since the data link is prone to be interrupted when one peer moves on. Therefore, researchers tend to focus on a novel architecture to guarantee the transfers via P2P WiFi.

A promising fact is that proxies can be employed in WiFi spots to prefetch content for mobile nodes such as vehicular offloading strategies. Moreover, a Subscribe-and-Send architecture was proposed in [139] to address this issue. In such an architecture, data offloading utilizes WiFi Direct D2D links. The main idea of WiFi Direct D2D links is to allow direct communications between peers instead of taking a detour through a traditional WiFi infrastructure. Each smart-phone running certain apps is considered as one node that is responsible for managing subscription and delivery. As shown in Fig. 10, the Subscribe-and-Send architecture is mainly based on the subscription table of the content service provider (CSP). The CSP records every subscriber's information including subscribed content name, subscriber's ID, and the deadline of the subscription. Once the contents are still not downloaded from peers until a deadline, the subscribers either download the subscribed contents or update the subscription deadline. This will allow subscriber to go on waiting for transfers via the WiFi connection. The subscription information will be removed from the subscription table on CSP when subscribers missed the deadline of P2P offloading or subscribed contents were successfully offloaded.



Subscription Table on CSP
Content1,Node_1,Deadline_1
Content1,Node_5,Deadline_5
Content1,Node_2,Deadline_2
Content1,Node_3,Deadline_3
…

Subscription Table on node
Content1,Node_1,Deadline_1
Content1,Node_5,Deadline_5
Content1,Node_2,Deadline_2
Content1,Node_3,Deadline_3
…

Fig. 10. A Subscribe-and-Send architecture [139]. Source node 1 is the node that is able to download the content from servers via cellular networks. Then it will check whether the content it downloaded has been subscribed by others by detecting the subscription tables on CSP. The content will be transmitted to the subscriber via relays by opportunistic P2P communications. Each peer on the route will exchange subscription tables with each other when they meet via WiFi networks.

It is notable that relays between WiFi nodes and P2P transmissions are based on opportunistic WiFi protocols as the key technology of P2P offloading. In the Subscribe-and-Send architecture, it is assumed that there are some users who are able to pay for bulk data transfers. On the other hand, they may prefer to download the content from the CSP via cellular network source nodes. First, subscribers send the request to the CSP to download the content from the CSP. Then, the subscription information is recorded in the subscription tables on both CSP and user's devices. Then, the source node that has downloaded the subscribed contents checks the subscription table on the CSP to probe whether the content it has downloaded is subscribed by others or not. Then, it will utilize users' IDs to deliver the content to subscribers one by one via opportunistic WiFi protocols. On the other hand, a user node can exchange its subscription tables with another node if they connect via WiFi networks. The subscription tables on users' devices are exchanged to probe whether one of the devices has the content that the others have subscribed or not. Then, the content will be transferred via WiFi networks once the information is matched. Thus, messages will be sent to the CSP when the content subscribed is successfully transferred via WiFi networks. Then, the subscriptions in those tables are removed. In order to assure the accuracy of subscription tables for all the nodes, the subscription tables on every node will be checked every 10 minutes by detecting subscription tables on the CSP.

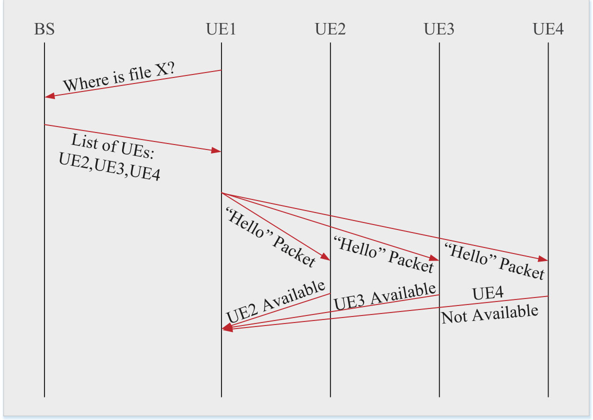


Fig. 11. Signalling message sequence for setup of communication between M2M UEs in cooperative networks [144].

It is worth noting that these subscription tables can be implemented by various opportunistic routing protocols in the routing layer [140], [141]. On the other hand, a High PRobability Opportunistic (HPRO) was proposed in [139] for the Subscribe-and-Send architecture as a novel opportunistic forwarding protocol. The basic idea in this protocol is that the contents are transferred only when the number of contacts reaches a certain threshold. This is to ensure that a considerable probability of successful delivering between peers. Moreover, the threshold is alterable along with the density of nodes and updated periodically to reply to changes of nodes' mobility status. So in the process of HPRO, the first step of a source node is to probe the number of contacts, then compare the number with the threshold. The subscribed content begins to be transferred only when the number is larger than or equal to the threshold. Otherwise this node waits for the next comparison while the parameters may change or be updated. Simulations and comparisons showed that the Subscribe-and-Send architecture with HPRO has higher delivery performance than several other approaches while more cellular network traffic can be saved.

3) Cooperative Relaying: Conventional cellular networks can provide information about active users in range of several tens of meters to each new UE to determine the potential peers of the group. Then these peers transmit the popular data via P2P communications with the help of already existing infrastructure of cellular networks. Other than complementary networks with multi-mode mobile terminals, it is essential to investigate cooperative networks of integrating P2P techniques by using a unified interface. Fitzek *et al.* [143] proposed an associated cooperative framework to dynamically combine cellular networks with P2P/short-range links. In this framework, UEs firstly form a cooperative group then each UE can exchange their received substreams over the short-range link including WiFi and Bluetooth. As shown in Table XII, the challenge is to design a common air interface to improve spectrum efficiency and reduce computational complexity. This common air interface focuses on splitting the spectrum for cellular networks and shot-range communication, and SDN might be a promising solution for implementation of the common air interface.

On the other hand, since the frequency spectrum with cellular networks is typically unused in the uplink, it is essential to utilize this unused spectrum to improve the system performance including the number of simultaneously served users, throughput, and QoS. Popova *et al.* [144] proposed a cooperative network architecture for efficient distribution of popular non-real time data content. Their cooperative networks utilize the available frequency spectrum from a cellular network such as UMTS (Universal Mobile Telecommunication Systems). Considering the different traffic distribution of data services between the uplink and downlink, this architecture exploited the normally unused uplink frequency band for a group-based cooperative M2M data dissemination. The paired spectrum allocation principle of UMTS FDD (Frequency Division Duplex) is used to balance data traffic in this architecture. To realize a direct M2M cooperative data exchange on cellular uplink bands, this proposed architecture integrated P2P techniques into the existing cellular structure. Typically, the original popular content is divided into several logical packets which are distributed among UEs. As shown in Fig. 11, UE1 first sends a request to the BS to obtain the information of potential UEs (e.g., list of UEs), where UEs are considered as servers, and then UE1 establishes the M2M communication link on its own currently unused uplink channel to download related packets from UE2, UE3, etc. Simulation results show that the overall downlink throughput gain can be considerably achieved by this M2M cooperative solution. They argued that combining the cooperative solution with cellular networks is a promising candidate for distribution of content in cellular networks. Moreover, integrating P2P networks with centralized controlled cellular networks can support a large range of wireless services. However, properly designed scheduling of the packet transfer, which significantly affects download time and system throughput of the content distribution, is not well addressed.

Energy consumption on UEs for multicast-based cooperative D2D has been studied by several researchers. First, BSs distribute common content to a group of UEs on long-range channels. Then UEs multicast the received content to other UEs via multi-hop cooperation (D2D communication). However, how to determine the energy consumption for this cooperation on UEs with fairness among UEs is still a challenge. Further research on optimal offloading of cellular networks need to be addressed, and reducing the required number of long-range channels is a promising approach. Al-Kanj *et al.* [145] optimized the chunk distribution and multicast transmission between the UEs and multi-hop cooperation in terms of constraints on the energy consumption of UEs. They formulated a mixed integer linear programming solution, taking into account the dynamics of the network. Since these problems are NP-complete, they have to present polynomial time greedy algorithms to derive computationally fast solutions. Simulation results show that significant cellular offloading gains can be obtained at the cost of a very small fraction of UEs' battery levels consumed on multicast.

TABLE XII
ISSUES OF COOPERATIVE RELAY MECHANISM

Issues	Challenges	Researchers
Combine cellular networks with P2P/short-range links	Design a common air interface	Fitzek *et al.* [143]
Exploit the normally underused uplink frequency band	Design a schedule of the packet transfer	Popova *et al.* [144]
Energy consumption on UEs for multicast-based cooperative D2D	Determine the energy consumption for multicast on UEs	Al-Kanj *et al.* [145]
Overlay multicast based on bi-directional TCP connection	Address a new reliable data delivery scheme for multicast	Kim *et al.* [146]
Next-generation service overlay network (NGSON)	Integrate service control functions with P2P overlay	Lee *et al.* [149]

IP multicast does not work well in the current Internet which is based on the unicast communications. Various methods have been addressed to solve this problem and one promising issue is overlay multicast. The main idea of overlay multicast is that bi-directional TCP connection (instead of TCP) is used between end-systems, where TCP may cause data loss and duplicated data reception when the UE in the multicast tree performs parent switching. Therefore, further studies on a new reliable data delivery scheme for the multicast communication need to be addressed. Kim *et al.* [146] integrated simultaneous communications between multiple senders and multiple receivers with n-plex multicast service [147] and devised a reliable data delivery mechanism. However, this overlay multicast does not support context awareness, dynamic adaptation, and self-organization of service overlay network (SON). Several studies have attempted to integrate service-oriented architecture with SON infrastructure. Nevertheless, it is difficult for SON to deal with ubiquitous and dynamic environment of UEs and services. Thus, next-generation SON (NGSON) was proposed by the IEEE P1903 Working Group (WG) [148]. Based on the service publication information from providers and UEs' requirements, NGSON can well control the interactions among distributed services by service control functions and service delivery functions. Service control functions manage the service information from service providers and help UEs to discover service instance that meets users' requirements. Then, service delivery functions enforce corresponding QoS control mechanisms. Lee and Kang [149] briefly reviewed the features of NGSON and showed that NGSON functionalities are practical and efficient to handle the future trends of dynamic environments. Meanwhile, further studies on content delivery support need to be considered. Moreover, it is a challenge for future researchers to integrate service control functions with P2P overlay to address the content delivery issues.

C. Multipath Transmitting

1) Multipath TCP Offloading: A crucial challenge for current offloading strategies is that continuous connectivity is not guaranteed by the TCP. This is due to the fact that the connections are prone to be interrupted once the status of the current interface changes with the TCP. There are many researchers focusing on devising novel protocols to extend the TCP to address this issue of interruption avoidance. Recently, several protocols were proposed utilizing a promising idea of exploiting multiple interfaces in [150] and [151].

An alternative approach to maintain continuous connectivity is to allow multiple paths via several networks to simultaneously transmit the data of one content. This is still valid for the scenario when the status of one interface changes with the interruption of an ongoing transmission. On the contrary, all approaches in the previous discussions are single-path schemes which only utilize one interface and network to transmit traffic by the TCP. However, the multi-path cannot be implemented by the TCP. As an extension to TCP working on the Internet, Multipath TCP (MPTCP) is devised as a multipath protocol in the transport layer to extend TCP to address this issue [152], [153]. MPTCP is a protocol responsible for managing several simultaneous connections via multiple networks.

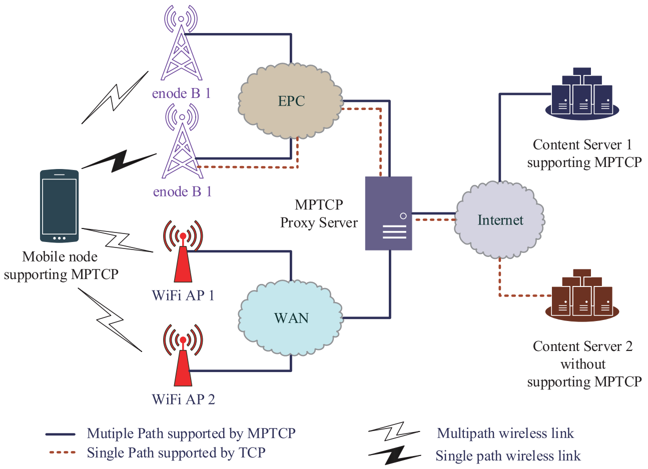


Fig. 12. Typical architecture of Multipath TCP.

As shown in Fig. 12, a mobile node equipped with MPTCP can download contents from remote servers via multipath. When the remote server is equipped with MPTCP such as content server 1 in this figure, the MPTCP proxy server just acts as a relay between the multipath connections established by both the mobile node and the content server 1. When the remote server is not equipped with MPTCP such as that of content server 2, the mobile node just establishes multipath connections with MPTCP proxy server. At the same time, the MPTCP proxy server communicates with the remote content server by TCP since content server 2 does not support the multipath communication. Therefore, MPTCP proxy server can significantly support mobility of users by multipath connections whether remote content servers are capable of MPTCP or not.

MPTCP can be deployed in end-hosts in a typical deployment strategy to avoid additional requirements on the network. This can be implemented by traditional infrastructures capable of controlling end-to-end communications. In addition to typical deployment, a MPTCP-enabled node is added on the path between end-hosts as a proxy, so traffic can still be transferred through multiple networks even when one end-host is not equipped with MPTCP. This is appropriate for the scenario where all the networks and proxies are managed by one operator. However, this scenario needs improvement of capacity in the backhaul when both WiFi and cellular networks are managed by the same operator. On the contrary, proxies should be deployed by a third party when WiFi and cellular networks are managed by several different operators. In such a case, there is no need for operators to enhance their backhaul capacity to meet the traffic increase on base stations, which provides extra operators' savings.

MPTCP can benefit from WiFi offloading from several aspects. First, there is a congestion control mechanism in MPTCP to balance traffic on each path to avoid congestion [154]. Second, evaluations provided that more than 50% of the channel holding time can be saved via WiFi by MPTCP, while high holding time may lead to congestions [155]. Third, the essential advantage of MPTCP is to support seamless mobility between multiple interfaces since multiple networks are utilized so a handover between networks cannot interrupt the ongoing session. Fourth, there is no need for updating the hardware to implement MPTCP. However, as far as energy efficiency is concerned, only transmitting data blocks can show satisfied performance unless a better activation controlling scheme is devised to determine when to utilize MPTCP. It is apparent that multiple concurrent links incur extra energy consumption. Chen *et al.* [90] made an attempt to propose a novel energy-aware MPTCP-based content delivery scheme (eMPTCP). This scheme balances support for throughput with energy consumption awareness. Their simulation shows that 14% gains of energy efficiency over MPTCP can be achieved with eMPTCP by offloading traffic from the more energy-consuming interfaces to others. Nevertheless, there is still large room for improving energy efficiency of MPTCP. Moreover, MPTCP is still in the draft state by IETF [156]. Further work is needed to make an attempt to enhance energy saving mechanism.

2) Concurrent Connections: The majority of the literature focus on utilizing WiFi AP and UE to dynamically switch interfaces in the cellular-WiFi network. However, the forthcoming devices are able to operate multiple wireless interfaces simultaneously. It is interesting to exploit several concurrent connections to transmit data flow to achieve various gains such as seamless experience. The server-based MPTCP mechanism extends traditional TCP to provide a transmitting protocol for multipath communication in the transport layer. Another mechanism IFOM enable UEs to concurrently maintain connections with the cellular network and a WiFi AP. While the main challenge is to reduce energy consumption, just as discussed in Section IV above. In addition, to support using existing cellular and WiFi links simultaneously, it is still interesting to investigate traffic splitting mechanism and evaluate its energy consumption. Abbas *et al.* [95] proposed a multi-incentive approach for traffic splitting. They formulate the tradeoff between throughput maximization and battery energy minimization into a function of the ratio of data to be split. They use experimental measurements with an Android device to evaluate their traffic splitting scheme, in terms of tradeoff between throughput and energy saving. Nevertheless, further work is needed to investigate energy saving to extend the battery time of UEs.

D. Collaborative Offloading With Delay

Owning to a successful offloading solution, cellular network operators can address the problem of rising traffic demand without significantly increasing their CAPEX and OPEX. On the other hand, WiFi access network providers need to get more revenue from customers and cellular operators. Therefore, a popular issue is how they cooperate so as to benefit both cellular network operators and WiFi access network operators.

As for user's benefits, a user can enjoy high bit rate if the user's network is allowed to handover to a WiFi network since the latter usually has a higher bit rate than cellular networks. In addition, a cellular operator always provides a coupon to their subscribers so as to encourage them to migrate their data to WiFi in order to leverage the load of the cellular network since the cost of the WiFi is lower than that of the cellular network. In general, a successful WiFi offloading approach should be beneficial to all participants including cellular operators, WiFi service providers, and end-users. Only in this way can participants accept and support it, which is significant to the development of a new approach.

Recently, a collaborative framework MADNET conforms to the idea mentioned above [157]. In this framework, cellular operators, WiFi service providers, and end-users can cooperate to make it feasible and effective. It is noteworthy that transmission performance in this approach can be improved a lot even in a sparse area of WiFi APs.

1) Opportunistic Offloading: In order to reduce the load of cellular traffic, cellular traffic offloading is very promising. An alternative approach is to use the distributive advantage provided by WLAN. The main idea is to send messages to hotspots which can be relayed to the aimed devices via WLAN so as to avoid using the data link through the cellular network. However, the precondition of this approach is that the WLAN is available for both, users and operators. That is since that the density of WiFi hot-spot in some remote areas may be low. Another fact is that the WiFi access networks are usually encrypted.

Another approach is "opportunistic traffic offloading" to overcome the above drawback. The main idea is to utilize the D2D connections such as WiFi direct and Bluetooth to transmit messages. When messages are transmitted to the partial devices, they relay the messages to the whole subscribers via inter-device connections without sending messages to every subscriber via cellular networks [158]. To better understand opportunistic traffic offloading, the difference between these two approaches is given in Fig. 13. The challenge is how to ensure the numbers of the partial devices that get the messages via cellular networks. However, in order to infer such a number, some essential information about mobile devices must be collected.

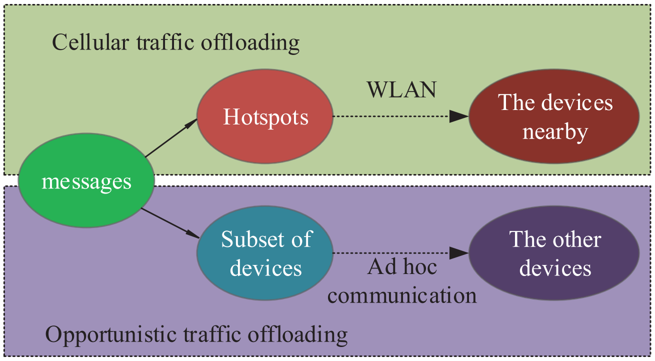


Fig. 13. Opportunistic traffic offloading compared to cellular traffic offloading. Unlike that the infrastructure is responsible for cellular traffic offloading, it is the subset of devices that relay the traffic by Ad hoc communications.

TABLE XIII
THREE DIFFERENT TOMP COVERAGE METRICS

Coverage Metrics	Factors Considered
Static coverage	Node's current position
Free space coverage	Node's current position, Speed
Graph based coverage	Node's current position, Speed, Underlying road graph

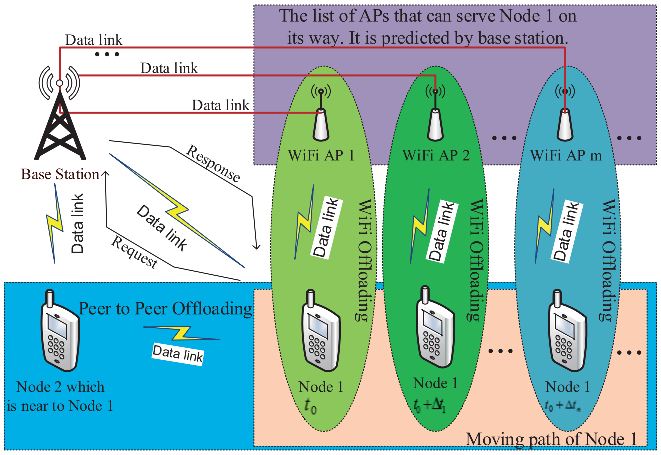


Fig. 14. The main idea of MADNET framework.

An alternative approach is to collect the historical information of social relations between the mobile devices that the messages should be transmitted to [159]. However, this idea is not easily feasible due to public concerns of privacy issues. With this in mind, the opportunistic traffic offloading using movement predictions (TOMP) was proposed in [160] to predict the inter-device connectivity. In this approach, the only information needed are the position and speed. This scheme is highly feasible with current cellular networks. In contrast to a typical cellular network, considerable energy can be saved as shown by the simulation results carried out for TOMP's proposed approach. In more details, as shown in Table XIII, there are three different coverage metrics in the TOMP's approach which include static coverage, free space coverage, and graph based coverage metrics. Those three coverage metrics vary in the number of cellular messages and delayed messages. However, the conclusion that can be drawn is that TOMP's approach can reduce the cellular traffic significantly without evident relays.

Another challenge for opportunistic offloading is high content delivery delays. High delays make opportunistic offloading not suitable for some small but timely content. Thilakarathna *et al.* [161], [162] proposed a hybrid content dissemination strategy in social mobile network scenarios. This strategy utilizes available networking infrastructure to replicate content on smart phones. Then this strategy leverages these replicators to propagate the content to others via opportunistic communication. The key challenge of this strategy is how to trade off between the content delivery performance and overheads of resource usage due to replication. To well address this issue, they make an attempt to ensure minimum content replication to reduce network infrastructure usage and devise an algorithm. Their trace-driven simulations show that replicating on just 10% of consumers significantly increases the content delivery to almost all consumers. However, their studies indeed do not make full use of social information to augmenting offloading scheme just like what Barbera *et al.* [163] and Rachuri *et al.* [164] have done. Further works are still needed to evaluate these offloading schemes' energy consumption for UEs and compare them to traditional delivery methods.

2) *Framework of MADNET:* The main idea of the MADNET framework is that it utilizes the cellular network to transmit some important content that has no delay tolerance. While it utilizes DTN techniques to transmit some content that is not urgent and has appropriate delay tolerance, or some content that is greatly large and leads to a higher cost for users. In particular, a cellular network is responsible for transmitting the request for content, while WiFi is responsible for delivering the block of data to mobile devices with delay tolerant techniques and opportunistic P2P offloading schemes.

As shown in Fig. 14, MADNET consists of cellular networks, WiFi networks, and P2P Pocket Switched Networks (PSN). There are four kinds of members in this framework including base station, mobile node 1, mobile node 2, and WiFi APs in the list. Node 1 is going to request offloading intention to the base station, and is moving on the route as shown in the figure. Node 2 is the peer that is near Node 1 and can be utilized to achieve offloading by opportunistic peer-to-peer offloading with Node 1. In a nutshell, Node 1 requests to a base station, then it offloads its traffic by WiFi links, either through WiFi AP in the list provided by a base station or Node 2 by a P2P WiFi connection. Thus, MADNET indeed adopts WiFi offloading and peer-to-peer opportunistic offloading which is applied for outdoor and mobile environments. In this framework, every node is able to not only upload and download content but also request the base station in range to help determine which AP or peer the content should be offloaded to. This decision is based on the fact that the cellular network can store the location of the neighbouring WiFi APs since they are fixed while users usually take a similar path. Therefore, it is easy to predict user's mobility patterns on the basis of the history of the stored data.

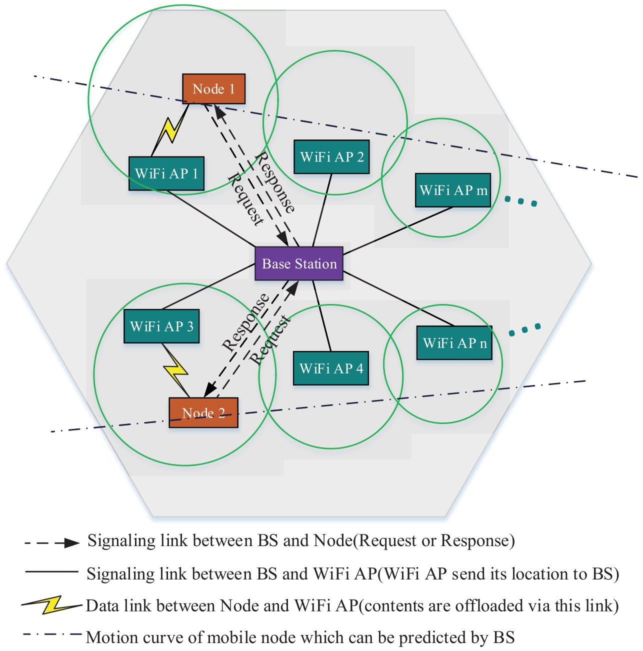

Node 1
WiFi AP 1
WiFi AP 2
WiFi AP m
Response
Request
Base Station
WiFi AP 3
Response
Request
WiFi AP 4
WiFi AP n
Node 2
Signaling link between BS and Node(Request or Response)
Signaling link between BS and WiFi AP(WiFi AP send its location to BS)
Data link between Node and WiFi AP(contents are offloaded via this link)
Motion curve of mobile node which can be predicted by BS

Fig. 15. The function flow of MADNET framework. First, mobile node 1 requests the base station and provides its speed, direction, and position information to the base station. Second, the base station predicts the route of node 1 and provides node 1 with the list of WiFi APs on the basis of the information provided by node 1. Then, node 1 begins to connect to the WiFi AP in the list with the highest priority and offloads its mobile data traffic through this WiFi AP.

Algorithm 1 Energy-Aware Offloading Decision Algorithm [89]

```
Require: The power of data reception $P_{3G}$ for 3G and $P_W$ for WiFi.
Require: The head and tail energy $E_T$ of 3G and $E_{oo}$ for the offloading related overhead.
Predict the throughput $B_{3G}$ for 3G network and estimate the offloading capacity $C_W$ and throughput $B_W$ of a WiFi AP.
Predict the prefetching capability $F$ of this WiFi AP.
Calculate the WiFi offloading duration $C_W/B_W$ and the time to receive the same amount of data through 3G network $C_W/B_{3G}$.
if $F \geq C_W$ and the following inequality holds
    $$E_T + P_{3G} \cdot C_W/B_{3G} - P_W \cdot C_W/B_W > k \cdot E_{oo} \qquad (2)$$
then
    Offload mobile data traffic to this WiFi AP.
else
    Repeat the above for other available APs.
end if
```

As discussed above, MADNET utilizes the cellular network as the signaling channel for controlling deliveries but utilizes WiFi network or peer-to-peer opportunistic offloading [159] as data link for transmitting the main part of content. MADNET selects several UEs as source nodes and then utilizes peer-to-peer opportunistic offloading to push the contents from source nodes to the rest of nodes via peer-to-peer connections. Then, the source nodes are responsible for delivering the contents to more users through short-range wireless connectivities such as Bluetooth and WiFi, etc. This mechanism can identify the social networks of the users then delivers specific contents to particular social groups. Thus, a large fraction of data can be offloaded from the cellular network based on the past history.

3) Procedure of MADNET: In the framework shown in Fig. 15, first, requesting node 1 provides its own status information that includes its position, speed and direction to the base station. Second, the base station predicts the route of the node and produces a list of APs that can serve the requesting node 1 on the basis of information that was provided earlier. Third, the list is sent back to node 1. Then node 1 will connect to the serving AP in line with the list. Finally, the APs in the list will deliver the content which should be transmitted by the cellular network.

A new WiFi AP with a greater RSS than the predefined threshold will be added into the list only when all APs are out of the range of the mobile node, and its priority level will increase once it is needed by one mobile node. In other words, the AP with the higher priority level has higher frequency of usage and is more requisite for users. In general, MADNET is an integration of WiFi and PSN combined with the cellular networks that are mature and have already been deployed widely by operators. MADNET benefits both users and operators in terms of the fact that users' cost of data delivery can be reduced and the problem of increasing data demand from users may be overcome. However, further integration of WiFi and opportunistic networks with cellular networks may still be a challenge in practice. More research is needed to be done for the strategic deployment of APs to minimize the number of APs without damaging users' satisfaction.

4) Energy Efficiency of MADNET: Without considering the energy efficiency, a WiFi offloading approach that tries to offload as far as possible may lead to battery depletion. As this question is considered, it is necessary for MADNET to increase energy efficiency. In order to reduce the energy consumption on mobile equipment, Ding *et al.* [89] proposed an energy-aware offloading decision algorithm for MADNET. This architecture aims to extend the battery life of a cellphone, namely, the mobile devices can reduce energy consumed and turn on the WiFi interface only when users request to offload.

As shown in Algorithm 1, the procedure of the energy aware algorithm is presented. This algorithm allows users to save energy, where P_{3G} represents the power of data reception for 3G, while P_W denotes the power of data reception for WiFi. The k is a parameter to accommodate measurement errors. In addition, the power state for 3G is unstable during state transition when UE activates 3G interface at the beginning or close to it at the end. E_T denotes such head and tail energy of 3G (energy consumed during state transition). Furthermore, offloading may cause extra energy consumption to get location information and to associate with the WiFi APs that are predicted to be available. E_{oo} denotes such extra energy consumption.

Thus, first, the base station in the cellular network predicts the throughput B_{3G} for 3G network, then estimates the offloading capacity C_W and throughput B_W of a WiFi AP. Second, the base station predicts the prefetching capability F of this AP. Third, the base station decides whether the offloading scheme should be initiated or not by computing the inequality in the algorithm. This inequality can be explained by that MADNET performs offloading only when receiving of prefetched data from WiFi APs saves more energy than the extra energy consumption E_{oo} as discussed above. Once determined, the base station will let the WiFi AP with high energy efficiency prefetch the content as far as possible. Then, the base station will predict the position and speed of users and send it back to the mobile device. Finally, the device will download the prefetched data from the previous WiFi AP. In conclusion, MADNET selects the most energy efficient APs for UEs. The selected AP which is responsible for prefetching data for UEs can well improve the utilization of WiFi channels to reduce the energy consumption on UEs.

Note that the offloading decision is affected by the binary opposition of the throughput of the WiFi and that of the cellular network. The data traffic will be migrated to WiFi if its throughput exceeds that of the cellular network. On the contrary, offloading will not be adopted when the throughput of the cellular network exceeds that of the WiFi network. However, the binary opposition of the throughput of two networks may vary with different scenarios. Thus, the decision needs to be confirmed by some measurement in these scenarios if necessary. Moreover, real-time network conditions may influence the performance of MADNET. In general, MADNET predicts the throughput of cellular networks on the basis of history and location information. Offloading may cause extra energy consumption when a mobile device sends its location information to the base station and communicates with APs. Further research studies are needed to eliminate the energy with the assistance of cellular networks by avoiding unnecessary connections on UEs with high mobility such as vehicular devices.

VI. Continuity in Vehicular Scenarios

Offloading data from cellular networks to WLAN for UEs indoor or some devices that are slowly moving has received much research attention already. At the same time, mobile data traffic from or to vehicle users is increasing dramatically. This is due to the huge increase of vehicles that contribute to vehicular communications using the cellular network. What is more interesting is that with the proliferation of Intelligent Transportation Systems (ITSs), it is more challenging to meet the data explosion from vehicles and be able to offload a part of data to different networks in an intelligent manner. On the other hand, both evolving cellular and short-range wireless modules will be part of future vehicles and will be added into mobile devices. However, unlike offloading motionless users, the different challenges come from two characteristics when dealing with highly mobile vehicles. The first characteristic is the high dynamics of vehicular communication environments, while the second is the fast fluctuating wireless channel, which needs proper handoff schemes or better transport protocols.

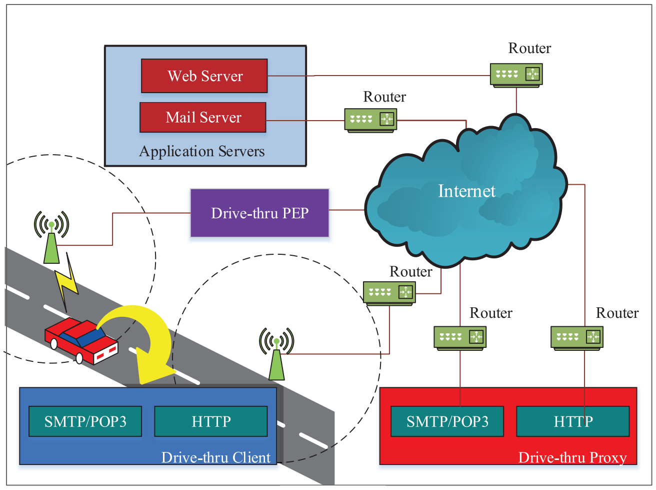


Fig. 16. Overview of Drive-Thru network.

A. WLAN for Vehicles

To solve the complex problem of intermittent feature of connectivity for vehicles with high mobility, as shown in Fig. 16, a Drive-thru Internet access was proposed [165]. There are two main components that include Drive-thru proxy and Drive-thru client in this framework. The main idea is to utilize those two entities as the intermediary to set up connections to relay and transport application sessions via peers in this architecture. This is done instead of establishing direct communications by clients and servers so that both of them can be used without frequent modifications. Therefore, the connection is relatively stabilized and capable of avoiding frequent changes with the environment. This is true since Drive-thru proxy and Drive-thru client are in charge of maintaining the persistent relationship initiated by vehicle clients on behalf of applications or servers.

As shown in Fig. 17, the responsibility of Drive-thru Client is to maintain persistent connections with its counter-part Drive-thru Proxy. Both of them implement the Persistent Connection Management Protocol (PCMP), which is a transport layer session protocol employed in both components [166]. The main function of the Drive-thru Client is to re-establish TCP connections in order to resume communication sessions whenever an interruption of connectivity is detected. To be specific, the Drive-thru Client acts as an application layer gateway at the application layer. The approach is to store requests of transmission from applications and cache them from its counter-part Drive-thru Proxy for the next connection during a disconnection period.

The responsibility of the Drive-thru Proxy as the counterpart of Drive-thru Client is to split direct connections between servers and applications. This can be achieved by concealing the characteristics of the network environment such as temporary state of disconnections of clients from servers to avoid unnecessary server's reactions. The main function of

TABLE XIV
COMPARISON OF TECHNIQUES REGARDING WLAN FOR VEHICLES

Techniques	Advantages	Disadvantages
Drive-thru Internet access [165][166]	Reduce TCP's reactivity Suitable for V2I communication	Hardly guarantee interactive data service Not suitable for V2V communication
Quick WiFi [167]	Speed up connection establishment by integrating all processes	Only suitable for low mobility scenarios
Cabernet transport protocol [167]	Well deal with intermittent connectivity and wireless losses	Only suitable for low mobility scenarios
IEEE 802.11 DCF based protocols [168]	Suitable for high mobility scenarios	MAC rate selection needs to be optimized

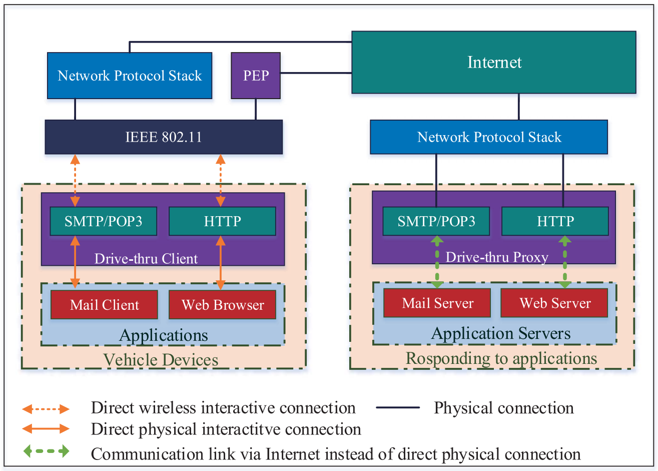


Fig. 17. DriveThru Internet Architecture.

Drive-thru Proxy is to relay data (for both mobile applications and servers) during connection periods. Then, it will buffer data from servers for mobile applications during disconnection periods. Moreover, the Drive-thru Performance Enhancing Proxy is an optional component to isolate the characteristics of WLAN to reduce the TCP's reactivity, and the characteristics are observed from the vehicle clients. As the feature of adjustment mechanism in TCP is considered, the Drive-thru Performance Enhancing Proxy is devised to avoid frequently changing states of the link layer.

However, similar to DTN, Drive-thru Internet access is not suitable for interactive or real time applications such as VoIP (Voice over Internet Protocol) and video streaming which are delay-sensitive. Actually, there is no design for some critical and urgent information to be transmitted in a proper way. In general, Drive-thru Internet is suitable for V2I (Vehicle-to-Infrastructure) communication but not for V2V (Vehicle-to-Vehicle) communication. In addition, as a single-hop WLAN system, Drive-thru Internet needs to improve the deployment density of hot-spots to ensure the quality of service, which leads to undesirable capital expenses.

There are still protocol issues attracting concerns for Drive-thru Internet as shown in Table XIV. First, as for reducing the connection establishment time, Quick WiFi was proposed in [167] to speed up connections. The main idea is that all related processes can be integrated into one process and the timeouts of related processes should be reduced. Simulations showed that the connection establishment can decrease to 400 ms. Second, to improve the transport protocol to deal with intermittent connectivity and wireless losses, Cabernet transport protocol (CTP) was also proposed in [167]. CTP utilizes a network-independent identifier to allow migrations between APs. It also sends probe packets periodically to distinguish wireless losses from congestion losses. Evaluation shows that CTP doubles the throughput of TCP. Third, in order to make the MAC protocols suitable for Drive-thru Internet in a high mobility scenario, an IEEE 802.11 DCF of large scale Drive-thru Internet was discussed in [168]. Such MAC protocols were built on the basis of IEEE 802.11 low mobility scenarios versions. Several other research attempts in the literature focus on the MAC rate selection and considerable level of goodput of TCP [168], [169]. But it is still a challenge to optimize the MAC rate selection scheme for Drive-thru Internet in the future.

B. *Offloading for Vehicles*

As far as the Drive-thru Internet feature is concerned, non-critical information can be offloaded via Drive-thru Internet. However, there are several challenges such as handoff strategies to reduce the time of establishing connections with WiFi including the association, authentication, and configuration time. Traditional hard handoff allows WiFi devices to handover to another WiFi AP with the highest RSS only when they are disconnected. This is due to the fact that devices may lose the chance to connect to strong WiFi APs during offloading with dissatisfied transmission performance. In general terms, probing the availability of WiFi may be a solution. However, the active scanning of WiFi availability while connected [170] is not energy efficient but results in delays for connection establishment.

In the Drive-thru Internet scenario, the handoff scheme has several aspects need to be improved. As for the throughput issue as shown in Table XV, a handoff in vehicular multi-tier multi-hop mesh networks was proposed in [170]. The main idea is that the data traffic is routed to APs with the backhaul connection via other APs (without backhaul connection) in a single or multi hop path. Evaluation showed that triple throughput can be achieved by this handoff scheme compared to the traditional handoff. Variational parameters of vehicular scenarios may lead to frequent disconnection of WiFi access and severely impact the performance of interactive applications. Thus, ViFi (V-band WiFi) was proposed in [171] as an optimal handoff scheme to address this predicament. ViFi utilizes the BSs in the vicinity of vehicles to relay packets in order to decrease the frequency of disruptions. This is possible since vehicles may suffer high bursty losses when using only one BS. Evaluation showed that considerable improvement of both TCP performance and VoIP services can be achieved by ViFi.

TABLE XV
RELATED CANDIDATES OF IMPROVING HANDOFF REGARDING VEHICULAR OFFLOADING

Issues	Candidates	Challenges
Improve throughput	Multi-tier/multi-hop based Handoff[170]	Traffic routing
Deal with disconnection	ViFi [171]	Relay techniques
Speed up handoff process	Scripted Handoff [173]	Prediction of vehicles' mobility
Prefetch	Cellular-based prefetch techniques [174]	Determine the relevant APs and prefetching part of data
Against packet loss	Prediction/Prefetch-based method [173]	Sensitive to the accuracy of predictions of mobility and connectivity

Another essential parameter of great concern is the time of the handoff process. It is important to reduce this time, especially when vehicles suffer frequent disconnections. This can be realized by reducing the time of availability probing and IP configuration via dynamic host configuration protocol (DHCP) [172]. In order to reduce the time of IP configuration, auto-configuration and cooperation between mobile clients and APs are utilized to avoid IP address assignment via DHCP when clients connect to APs.

In order to reduce the time of availability probing, the information (such as network name, MAC address) about road side units (RSU) is ought to be restored and cached for the next quick connection establishment when vehicles are driving on familiar routes. This is realized by the prediction of the WiFi connectivity and the mobility of vehicles. The historical information of vehicles that can be collected to predict the connectivity and mobility of vehicles includes GPS track, RF fingerprint, link and network address of APs. After the collection process, estimate of connectivity of APs can be derived. Both, vehicles' mobility and APs' connectivity, are used as inputs to compute the locations where vehicles should handoff and which AP they should connect to during periods of inactivity. This was named Scripted Handoff in [173].

On the other hand, transferring strategies for vehicular offloading need to be resolved because of lossy link performance and frequent disconnections. Prefetch is a promising approach to improve download performance (throughput per hour, the number of disconnections per hour). The APs cooperate with vehicles so as to determine the part of content that vehicles are going to download. Then, APs prefetch the part of content from servers to prepare a distributed cache for vehicles. A technique was proposed in [174] that utilized cellular networks to transmit the prefetch request to the prefetch units embedded in relevant APs, which are responsible for prefething data from the Internet. Vehicles then receive the prefetched data directly from APs instead of downloading the prefetched data from a server when vehicles are connected to the relevant APs. So the key is to determine the set of relevant APs where vehicles should be connected.

To address the handoff and prefetch issues, a predictive method based on historical information on familiar routes was proposed in [173], including predictions of WiFi connectivity and vehicles' mobility. Based on such predictions, a specialized transport protocol for prefetching was designed without being sensitive to packet loss. This protocol is however sensitive to the accuracy of predictions of mobility and connectivity that are based on several assumptions. The mobility prediction is based on the assumption that people's driving habits are highly predicable. The connectivity prediction is based on the assumption that the connectivity information including physical layer information, link layer information, and network layer information are stable for a long time. Evaluation showed that the accuracy of mobility estimate is bad for signaled highway while excellent for expressway and village roads. More analysis showed that information such as RF fingerprint and SNR remains moderately stable over several weeks, which is useful for prediction. For all of that, more accurate prediction mechanisms need to be designed to optimize offloading performance.

C. Opportunistic Offloading via Multi-Hop Networks

It is important to better utilize the advantages of both cellular and WiFi access networks and assure both critical and non-critical information be transmitted in a proper method. A good approach is that a cellular network is responsible for maintaining always-on connectivity to transmit critical, delay-sensitive, and real-time information for vehicles, while WLAN is responsible for offloading non-critical, delay-non-sensitive, and non-real-time information. Therefore, a mechanism that implements enhanced cooperation and integration of cellular networks for vehicles is necessary.

However, the challenge is to determine a proper WLAN system for vehicles to offload traffic. Unlike the single-hop network access of Drive-thru Internet, a promising WLAN for vehicular ad hoc networks (VANETs) is the multi-hop wireless network with a short range [175]. Composed by vehicles in a certain range and infrastructure deployed along a road, VANET supports both vehicle-to-vehicle and vehicle-to-infrastructure communications. Therefore, mobile data traffic can be offloaded to WLAN through both direct links to RSUs and Multi-hop links of VANETs.

Although multi-hop wireless networks provide a solution of offloading for vehicles, there are still several challenges to be solved. In most cases, only a part of traffic of vehicles can be offloaded to WLAN, instead of the total traffic. This is due to the fact that the connectivity is precarious and dynamic as the intermittent feature of WLAN for vehicles. So the challenge is how to decide which part of the traffic should be offloaded to the WLAN. On the other hand, in order to alleviate the affect of handover on user experience, a handoff mechanism for vehicles should be able to migrate the traffic seamlessly. A promising approach is IP flow mobility which aims to migrate IP flows selectively between interfaces via multi-hop VANETs. IP flow mobility sends addressed packets via different routes to the same destination. This multi-path scenario has been standardized by the IETF, as discussed in Section V-C.

However, upon surveying the literature, we have concluded that not much has been done on IP mobility flow, seamless handover or multi-hop WLANs connectivity to offload traffic opportunistically. In order to better utilize multi-hop feature of VANETs, a Seamless Internet 3G and Opportunistic WLAN Vehicular Internet Connectivity (SILVIO) was proposed [176]. SILVIO utilizes a tree-based routing protocol [177] as the routing protocol in VANET to solve the IP address configuration and routing issues. Based on the prediction when the interface connected will disappear or emerge, the Media independent handover (MIH) (IEEE 802.21) [178] was adopted in a SILVIO in order to optimize the handover schemes and minimize the loss of packets. It assists operators to choose their own policy by adding media independent handover function. It is also capable of providing seamless experience while moving across different ranges of wireless interfaces.

There are four handoff schemes based on how eager the vehicles are offloaded to VANET. The first one is SILVIO direct, which utilizes VANET to offload when vehicles are able to connect to RSU directly as a single-hop path. The second one is SILVIO conservative, which utilizes VANET to offload only when the number of hops on the path to RSU is less than a threshold so as to limit overmuch relays and delays. Vehicles will back off to the cellular network if the connection is interrupted but will not connect to the RSU once again in the same area in order to avoid the ping-pong effects even if a new WLAN interface is detected. The third one is SILVIO persistent, which is different from the second for that vehicles will choose to connect to the RSU more than one time while less than a proper number of times. The third one is only suitable for the case where vehicles are moving closer to a RSU. The fourth kind is SILVIO sticky, which is different from the third scheme for that vehicles will still connect to a RSU to offload data even when vehicles are moving away from the RSU.

Simulations in [176] showed that on average, the SILVIO sticky solution performs better than that of the SILVIO persistent. This is due to the fact that since vehicles can still utilize the RSU to offload traffic even when they are getting away from the connected RSU. However, in this solution, pairing RSUs of opposite directions for coming/leaving vehicles or vehicles should be deployed along the road to avoid unfairness. The unfairness may occur when vehicles from opposite directions may utilize the same RSU simultaneously. This unfairness can lead to economic challenges to operators. Thus, the SILVIO persistent is the best choice as far as economic restrictions and deployment feasibility are concerned. All in all, evaluation showed that the connection time (when vehicles connect to a VANET via multihop paths) can be improved by 80% if SILVIO is utilized.

As far as features and environments of vehicular offloading are concerned, SILVIO is a promising solution since it is based on existing standards. Both vehicles and operators benefit from opportunistic offloading via multi-hops of VANET. That is true since vehicles can enjoy a wide bandwidth at a low cost while they can still enjoy the always-on cellular connections to exchange critical information and get Internet service timely. In addition, although higher offloading efficiency than previous efforts in a metropolitan area can be achieved by a deployment algorithm based on density of users' request frequency in [179], more parameters and practical factors need to be considered to achieve an optimized offloading ratio. Therefore, multi-hop communication and opportunistic peer-to-peer schemes need further studies to improve the vehicular offloading effectiveness. For instance, further improvements may be achieved by optimizing deployment of APs in an advanced approach. Finally, a summary of related challenges is shown in Table XVI.

TABLE XVI
SUMMARY OF SILVIO [176]

Items	Context
Techniques	1. A tree-based routing protocol [177] 2. Media independent handover (MIH) [178]
Advantages	1. Well assist operators to choose their own policies 2. Provide well seamless experience
Challenges	1. Consider more parameters and practical factors 2. Investigate advanced deployment strategy of APs

VII. Future Research Directions and Challenges

A. Comprehensive Classification

In this section, as shown in Table XVII, we provide a comprehensive guidance in the field of WiFi offloading schemes. To the best of our knowledge, it is the first work to classify state-of-the-art of WiFi offloading schemes by combing the incentive techniques for WiFi offloading. To help researchers quickly find a specific study incentive in their forthcoming research in this area, we first classify state-of-the-art WiFi offloading into five categories in the horizontal direction of Table XVII, taking into account their different incentives. These five categories consist of capacity, cost, energy, rate, and continuity. We have discussed these incentives one by one in previous sections.

Furthermore, for the convenience of figuring out which main technique is adopted in various works for readers, we further classify the state-of-the-art into five categories in the vertical direction in this table, considering whether WiFi offloading schemes adopt these five key techniques.

1) Five Key Techniques: We respectively summarize the characteristics and challenges of the five techniques as follows.

a) *Non Delay/Delay-Tolerant:* If a WiFi offloading scheme does not distinguish delay tolerant data from the data traffic being offloaded, it is regarded as non-delayed offloading. In fact, it is a smart approach to utilize the characteristic of delay-tolerant data services such as multimedia downloading, which is regarded as delayed offloading. The challenge is to investigate how a good performance of tradeoff between delay, throughput, cost, energy, and rate can be achieved, in terms of various scenarios or requirements from users and practical environments. To the best of our knowledge, most attempts are made to consider the tradeoff between just two factors (delay and throughput). For various incentives,

TABLE XVII
A NEW WIFI OFFLOADING CLASSIFICATION

Incentive Technique	Capacity	Cost	Energy	Rate	Continuity
Non-delayed	[38][39][42][44] [24][25][46][47][48] [49][50]	[70][63][69][72][66][73] [67][68][75][76]	[83][89][90][85][80] [86][81][93][91][94] [95][97][96][92]	[102][103]	[125][126][127] [139][139][129][130][131] [150][151][152][153][154]
Delayed	[37][52][53][51][54] [55]	[64][65][71][77][78]	[88][84]	[98][100][101][6]	[158][160][159][89]
Non-Cellular	[37][44][24][46] [48][52][53][55]	[63][77][72][78][73] [75][76]	[83][90][91][95][97] [96][87][92]	[102][103]	[125][126][127]
Cellular	[51][54][38][39][25] [47][49][50][51][54]	[64][65][71][70][72] [66][67][68]	[88][89][84][85][80] [86][81][93][94]	[98][100][101][6]	[129][130][131][134][139] [139][150][151][152][153] [154][155][158][160][159]
Non-D2D	[50][51][37][54][38] [39][44][24][46] [47][48][49][54][51]	[64][65][71][70][63] [77][72][78][66][73] [67][68][75][76]	[83][89][84][90][81] [93][91][94][97][96] [87][92]	[98][102][103][100][101] [6]	[125][126][127] [150][151][152][153][154] [155]
D2D	[50][51]		[88][86]		[129][130][131][134][139] [139][158][160][159][89]
Non-Prediction	[51][37][54][38][39] [44][24][46][49][53] [51][54]	[71][70][63][72][78] [66][73][67][68][75] [76]	[83][85][80][86][81] [93][91][94][97][96] [92]	[98][102][103][100][101] [6]	[125][126][127] [129][130][131][134][139] [139][150][151][152]
Prediction	[50][47][48][50] [52][55]	[64][65][77]	[88][89][84][90][87]		[158][160][159][89]
Non-Prefetch	[51][37][54][38] [39][44][24][48][49] [52][53][51][54][55]	[64][65][71][70][63] [77][72][66][67][68] [75][76]	[88][83][90][85][80] [81][93][91][94][95] [97][96][87][92]	[98][102][103][100][101] [6]	[125][126][127] [129][130][131][134][139] [139][150][151][152][153]
Prefetch	[50]	[78][74]	[89][84]		[165][167][170][171][173] [175][176]

there is room for utilizing delayed offloading to augment current WiFi offloading schemes.

b) *Non Cellular/Cellular-Assisted:* If a WiFi offloading scheme just focuses on enhancing the network selection scheme on the UE side to switch interfaces during offloading, it is regarded as a non cellular offloading. In fact, from a comprehensive perspective in the total network, cellular network can help transmit controlling signals of WiFi offloading schemes. Moreover, BSs in a cellular network are responsible for controlling the offloading procedure. This is regarded as cellular-assisted WiFi offloading. To realize a smart offloading or an automatic offloading, a cellular network usually has to collect information from WiFi APs and UEs simultaneously. The challenge is to devise a mechanism to achieve the best performance with the smallest information. Most mechanisms focus on utilizing as much information from UEs as possible, ignoring extra congestion and privacy issues. Further work is needed to investigate more intelligent prediction mechanisms to help cellular BSs perform WiFi offloading.

c) *Infrastructure WiFi/D2D-Based:* If a WiFi offloading scheme just utilize infrastructure-terminal WiFi links to offload data traffic, it is regarded as Infrastructure WiFi offloading. A con is that these schemes are very dependent on high density of deployment of available WiFi APs. However, an increasing number of works utilize ad hoc network to augmenting infrastructure WiFi offloading. Note that some close-by mobile nodes may have the same demand for data services, it is very interesting to utilize D2D links to share these contents that one mobile node has cached. The challenge is to investigate how UE probe cached contents in vicinity that user demand and then perform transmission via multi-hop links. It is more interesting that a UE request the mobile node in vicinity which can access available WiFi APs to download contents, even it has not cached before.

d) *Non Prediction/Prediction-Based:* To make a good decision for WiFi offloading, some offloading schemes usually utilize historical information or capture some instantaneous information about WiFi APs, cellular networks, and UEs to perform a decision function in their algorithms. This scheme is regarded as a non-prediction offloading scheme. Actually, for a smarter offloading scheme, it may utilize some essential information to predict UEs' route and potential throughput. Then it performs an allocation mechanism to intelligently allocate resources to them, in terms of bandwidth or APs. The resource may refer to the best APs on their route, or bandwidth they need. This kind of scheme is regarded as prediction-based offloading. A main challenge is to find an appropriate mobility model in simulation for different schemes. Further work is needed to enhance their approximation.

e) *Non Prefetch/Prefetch-Based:* Most offloading schemes perform offloading procedure only when users request some data services. It is regarded as non-prefetch offloading. While it is interesting to utilize prefetch mechanism to augmenting offloading schemes. In addition to prefetch mechanism in MADNET, a fraction of state-of-the-art investigates this mechanism on UEs. UE may prefetch some data on its cache region when it accesses an available WiFi networks before user request this content. This scheme is regarded as prefetch-based offloading. However, it is still not clear how to avoid unnecessary prefetch and precisely predict users' potential data demand.

2) Future Directions of Five Techniques: We respectively provide the future directions of five techniques as follows.

a) *Non Delay/Delay-Tolerant:* Neither of them can solely show good performance. Future offloading is a developed integration of delay and non-delay offloading. The open issue is to find the optimal delay bound [55]. Most current studies make attempts to consider trade-off between just two factors (delay and throughput). Thus, a key future direction is to focus on more information including content type [42], users' mobility, users' behavior [39], [50], prediction of connectivity, and deployment of WiFi APs [41], [179]. Another direction is to implement the integration of non/non-delay offloading on current smart phones by utilizing a cloud cooperated heterogeneous networks [54] to control the offloading procedure with the help of cellular networks. Particularly, other than determining the specific content, deriving the specific portion of traffic delayed on UEs is also a future direction.

b) *Non Cellular/Cellular-assisted:* Current studies focus on utilizing cellular networks to transmit controlling messages, executing computation on BSs after collecting users' information [64], and providing cellular networks' information for UEs [77]. In the future, cellular networks are responsible for controlling the offloading procedures on BSs and UEs with a systemic controlling, instead of automatic controlling on one side and manual controlling on the other side. Particularly, the allocation mechanism of UEs/traffic should guarantee seamless experience. An open issue of this cellular-controlled network is to devise a mechanism to achieve the best performance with the smallest/incomplete information by considering users' privacy issues. Overall, the key future direction is to construct a congestion-aware and cross-system learning framework for femtocell/SCBS to avoid extra messages, or to investigate more intelligent prediction mechanisms based on instantaneous dynamics.

c) *Infrastructure WiFi/D2D-Based:* Since the quality of D2D connections is not guaranteed with high UE densities, it is essential to study the D2D-aware radio resource allocation and management mechanism in the future. For example, UE/Node with high transmit power can be provided a high priority for selection. Furthermore, for augmenting the quality of D2D connections, some methods including interference coordination/cancellation, advanced network-controlled management, efficient spectrum sharing, and device-grouping-based D2D-aided point-to-multipoint (multicasting) transmission schemes [180], [181] can be used. In the future, network bandwidths may be substantially underutilized for ultra-dense network scenarios. What is worse, dynamic systems have not been well studied as broadly as real dynamics. It is essential to focus on advanced real-time mechanisms considering user's variability, traffic, and dynamic systems [182]. Meanwhile, the complexity problem has to be well taken into account. In fact, it is not easy to trade off between complexity and dynamic systems. On the contrary, utilizing incomplete information based on different realistic scenarios is another interesting direction.

d) *Non Prediction/Prediction-Based:* For instantaneous functions (non-prediction) and prediction-based approaches, the future direction is the tradeoff between the computation complexity and real-time operations of offloading process for small cells. In the future, prediction-based offloading may require relatively less instantaneous inputs than non-prediction with the help of machine learning techniques. For prediction-based offloading, a future direction is to improve the accuracy of prediction for APs locations, connectivity and mobility of UEs [64], offloading potential (throughput), and delay tolerance. To better mathematically study a baseline integrated LTE-WiFi system, it is necessary to develop underlying computations for individual WiFi and LTE features [103]. Particularly, for context-aware offloading [77], a key direction for the prediction of WiFi connectivity is to extend the Markov models to a mixed model for user arrivals (e.g., integrate Poisson point process with clustered processes of user arrivals).

e) *Non Prefetch/Prefetch-Based:* On the whole, current studies regarding offloading pay few attention to prefetch techniques. In fact, there are some open issues including determining the popular content for offloading and deriving the best part of content for multi-hops/multi-path offloading within proxies and UEs. A direction for operators is to integrate prefetch and association mechanism with current cost-aware framework such as Win-coupon [64]. For example, if the requested delay-tolerant content is more "popular", related UEs should have a higher priority in the auction. An interesting direction might be to explore geographic information to help determine the "popular" content (e.g., based on data mining technology). Also, another interesting direction might be to explore advanced caching strategies to determine the best part of data (e.g., caching strategies based on information-centric networks and ad hoc networks).

In a nutshell, these classifications considering five incentives and five techniques provide a good guidance for researchers in the future.

TABLE XVIII
OPEN ISSUES SUMMARY

Incentives	Open Issues
Capacity	1. Augment on-the-spot offloading (instantaneous factors)
	2. Determine the best portion of traffic to be offloaded
	3. Investigate the impact of change of mobility
	4. Address deploying strategy of WiFi network (location)
	5. Tradeoff between delay deadline and offloading amount
	6. Find the best number of APs (formulating/measurement)
	7. Maximize the per user throughput by formulating
Cost	1. Maximize users' cost by cache and prefetch mechanism
	2. Integrate DTN case with Infrastructure case
	3. Investigate offloading market under incomplete information
Energy	1. Address seamless handover protocols translation
	2. Exploit automatic interface exchanging mechanism
	3. Offload via multiple wireless interfaces simultaneously
	4. Integrate multiple data streams with D2D techniques
	5. Save energy with mobility prediction and prefetching
Rate	1. Address congestion-aware network selection problem
	2. Desing intelligent and dynamic traffic steering mechanism
	3. Analyse on the load coupled network for WiFi offloading
	4. Investigate load balancing in MADNET
	5. Integrate dynamic data flow splitting with RATs
Continuity	1. Investigate real-time switch between multiple interfaces
	2. Improve context aware handoff mechanism
	3. Integrate relay/multi-hop techniques with offloading
	4. Address multipath transmitting protocols
	5. Augment opportunistic offloading and MADNET

TABLE XIX
FURTHER RESEARCH DIRECTIONS

Subjects	Further Directions
Standards	1. Augment multipath protocols with energy considerations (i.e. MPTCP and eMPTCP)
	2. Investigate seamless shifting for current sub-flow protocols (i.e. IP Flow Mobility (release-10 of 3GPP))
	3. Provide IP mobility with current Proxy Mobile IPv6 (Multiple concurrent data streams with D2D techniques)
	4. Simplify the certification and authentication procedure (Integrate Project Fi/ Passpoint techniques for offloading)
	5. Integrate intelligent connections protocols with offloading (CTP [167], DHCP [172], MIH(IEEE 802.21) [178], etc.)
Models	1. Exploit queueing analytic model & Cache mechanism
	2. Model hotspot deployment problem in realistic scenarios
	3. Model users' mobility and predict users' route
	4. Aggregate realistic models for different incentives
	5. Model interference mitigation
Methods	1. Address context-aware & Security and privacy issues
	2. Differentiate downlink and uplink in WiFi offloading
	3. Utilize load balance/load coupled mechanism
	4. Design high-capacity low-latency backhaul infrastructures
Others	1. Utilize beam combining techniques to determine the best directions for directional antennas
	2. Investigate programmable traffic control mechanism (SDN)
	3. Consider spectrum aggregation: LTE-Unlicensed (LTEU)
	4. Augment D2D-aware radio resource allocation/management

B. *Open Issues and Future Research Directions*

For the convenience of finding related open issues quickly, we have summarized and listed open issues (as discussed before) with different incentives (capacity, cost, energy, rate, and continuity) in Table XVIII.

Moreover, in addition to challenges we introduced above, we also list future research directions for WiFi offloading in Table XIX. For example, most of the studies focused on downlinks for WiFi offloading, few researchers [91], [94], [97] devote their works to investigating augmenting WiFi offloading within uplink scenarios. Considering that mobile content creation and uploading make uplink offloading a rising issue, it is still essential to design an intelligent offloading scheme by differentiating downlink and uplink. A recent study [91] shows that there is room for improvement in the uplink access scheme in terms of energy efficiency.

In addition, to reduce completion time of data transmission, it is significant to reduce the time when UE access WiFi APs. It is very interesting to simplify the certification or authentication procedure for WiFi offloading schemes. In fact, Google recently brought out a mobile carrier service called Project Fi that would compete with AT&T (T) and Verizon (VZ). Project Fi uses spectrum from T-Mobile and Sprint to create an affordable plan for UEs. It utilizes a network quality database to help UEs select high quality networks. It allows specific Android devices (e.g., Nexus6) [183] to automatically switch between cellular and WiFi networks. Project Fi supports seamlessly transition from one interface to the other. The trouble is that some other UEs have to update with the specialized SIM-level technology to support Project Fi [184].

On the other hand, in 2012, the WiFi-certified passpoint program had been promoted by WiFi Alliance, as an extension of the WiFi technology. Aruba Networks [185] also presented their WiFi certified passpoint architecture for public access in their white paper. Such extended WiFi can well enable seamless offloading for UEs. However, few works made some attempts to utilize it to augment WiFi offloading schemes. With an incentive of improving offloading performance in terms of capacity and energy saving, Hoteit *et al.* [186] compared Passpoint-based offloading with Passpoint-agnostic offloading. Their evaluations using real mobile data in Paris demonstrate that 15% capacity gain and 13% energy saving have been achieved with Passpoint-based offloading. Nevertheless, there is a large space for augmenting WiFi offloading schemes utilizing Passpoint techniques.

There are still other emerging issues related to WiFi offloading in current 5G/WiFi studies. As for capacity gains, recent studies [187] show that capacity for random pointing angles of directional antennas in the single best directions can be improved up to 20 times compared to today's fourth-generation Long Term Evolution networks. Meanwhile, this capacity improvement can be done with little increase in interference and required number of 5G base stations. However, further studies are essential to utilize beam combining techniques to help determine the best directions (strongest transmit and receive power) for directional antennas.

As for uncontrollable status for WiFi/cellular networks, it is essential to recover controllability and optimize traffic control mechanism. One interesting issue is to design a programmable traffic control scheme to direct traffic to the appropriate network by integrating SDN techniques. As for service continuity, LTE-Unlicensed (LTEU) is a promising way. LTEU extends LTE to unlicensed spectrum by a unified radio technology. LTEU can well provide seamless user experience. Further studies are essential to address interference modeling issues for LTE/WiFi coexistence over unlicensed bands. In addition, an incorporated consideration of intercell interference between macrocells and LTEU small cells are also essential [188].

In addition, further investigations related to dynamic use of the unlicensed band integrating cognitive radio network are also needed. The main idea of dynamic use of spectrum is sharing it in a way that allows spectrum bands unoccupied by primary UEs to be exploited opportunistically by secondary UEs. The basic rule is that primary UEs have the priority to use the spectrum while the secondary UEs can be preempted by primary users. Most of previous studies focus on the dynamic spectrum occupation process [189], [190], cooperative spectrum sensing scheme [191], spectrum and power allocation [192], [193]. Fox example, since improving the unlicensed spectrum utility from secondary UEs without interfering the licensed spectrum is a key issue, Hu and Zhu [189] investigated an underlay cognitive radio system to enhance access opportunities of the secondary users. More realistic and validated models for channel idleness as the foundation of credible cross-layer analysis are increasingly needed, Ghosh *et al.* [190] used two sets of real-time measurements to model channel idleness and build a predictive model by computing the availability probability of channels. Since spectrum sensing is of great importance for the secondary user to capture the under-utilized frequency bands, Nabil *et al.* [191] presented a channel assignment scheme for cooperative spectrum sensing. Simulations show that their scheme can well reach acceptable sensing probabfilty of error in terms of reduced sensing delay when compared to non-assignment. Lacatus *et al.* [192] presented a bit spectrum pattern-based wireless unlicensed system that successfully coexists with the licensed systems in the same spectrum range. Unlicensed system capacity for the optimal spectrum and power allocations can be considerably maximized in their work. However, white space spectrum in cellular offloading has not been fully investigated. Cui *et al.* [194] performed extensive spectral measurements of unlicensed bands to improve network load and proposed an algorithm to estimate the power consumption for the use of unlicensed bands in cellular offloading. Measurement-driven numerical simulation shows that networks with white space bands reduce the power consumption by up to 513% in sparse rural areas over WiFi-only solutions. Considering that FCC (Federal Communications Commission) allows unlicensed use of underutilized white space, a spectrum access system (SAS) is needed to implement an innovative spectrum management system. This system still requires more researches over a new end-to-end architecture, component protocols and novel radio systems. Although Kim *et al.* [195] presented their candidate architecture and its end-to-end implementation, further steps in this direction of new end-to-end architecture, component protocols and novel radio systems will be still needed in the future.

TABLE XX
EMERGING IDEAS AND STANDARDS FROM INTERNATIONAL ORGANIZATIONS

Organizations	**Emerging Ideas**
3GPP	1. Support dual connectivity transforming
	True Multipoint-to-Multipoint
	2. Improve spectral efficiency/Reduce Out-of-Band Emissions
	Filter Bank Multicarrier (FBMC) Waveform Design
IEEE	1. Address spectral localization issues: New waveform design
	Advanced multicarrier waveforms (FBMC) with low PAPR
	2. Spectrum extension
	Extend spectrum beyond 6 GHz
	3. Deployments of mmWave small cells
	Match the predicted growth rate of wireless traffic

C. *Emerging Ideas From Organizations*

Finally, to help researchers cover more emerging ideas of WiFi evolution, 5G, and related standards, we also summarize and list these emerging ideas in Table XX. In general, 3GPP standards evaluate from legacy centralized access structures towards heterogeneous access networks. Meanwhile, spectrum resource allocation needs to be enhanced to improve spectral utilization and efficiency. Thus, current Point-to-Multipoint (P2MP) and Multipoint-to-Point (MP2P) need to be evaluated into true Multipoint-to-Multipoint (MP2MP). In addition, it is promising for current OFDM (Orthogonal Frequency Division Multiplexing)-based LTE system to be evaluated into FBMC-based system. Filter bank multicarrier (FBMC) waveform design is a promising candidate for 5G physical layer and is being currently investigated by many other organizations [196]–[199]. FBMC is also an orthogonal multicarrier transmission scheme. Similar to 3GPP, it is essential for IEEE standards to investigate waveform design such as FBMC to address spectral localization issues. Especially, advanced multicarrier waveforms with low PAPR (peak-to-average power ratio) are crucial in terms of energy efficiency. Moreover, it is also important to extend the bandwidth beyond 6 GHz to consider higher bandwidth per wireless link and lower interference range. In addition, recent studies [200] show that it is crucial to let density of mmWave small cells match the predicted growth rate of wireless traffic with mmWave small cells. This prediction is based on modeling statistical behavior of user traffic services.

VIII. CONCLUSION

Future data explosion is expected to incur congestion in heterogeneous networks especially cellular networks. It is particularly important to deploy WiFi offloading schemes to shift some data traffic from cellular networks to WiFi. However, there are still some challenges that need to be addressed to augment WiFi offloading schemes with various incentives. An increasing number of research attempts focus on enhancing network selection schemes, optimizing cooperative mechanisms, determining the best fraction data being offloaded, allocating optimum APs to UEs, and so on. For the convenience of helping researchers find this particular research interesting and study it further, we classify the state-of-the-art in this field into five categories, considering different research topics that can be addressed under this category. We call this metric in our classification as incentive. These categories consist of capacity, cost, energy, rate, and continuity. In fact, capacity is the most basic incentive of WiFi offloading. That is also why WiFi offloading is needed in terms of the industry and academia. From operators' or users' perspective, we also investigate more offloading schemes according to the incentives of cost, energy and rate, respectively. Moreover, we identify a very important incentive referred to as continuity and extend common mobile scenarios to vehicular scenarios. We discuss these different offloading schemes, presenting their challenges, advantages, limitations, and further directions. Above all, we provide a guide for researchers by classifying state-of-the-art into further vertical five categories. A detailed discussion of open issues (Table XVIII), further related research directions (Table XIX) for augmenting WiFi offloading schemes, and emerging ideas and standards from organizations (Table XX) are also provided. This guidance can well help researchers to quickly find an interesting approach to achieve better performance for WiFi offloading.

References

[1] (Feb. 2016). *Cisco Visual Networking Index: Global Mobile Data Traffic Forecast Update 2015–2020 White Paper*. [Online]. Available: http://www.cisco.com/c/en/us/solutions/service-provider/visual-networking-index-vni/white-paper-C11-520862.html

[2] O. Cabral *et al.*, "Optimal load suitability based RAT selection for HSDPA and IEEE 802.11e," in *Proc. IEEE Wireless VITAE*, Aalborg, Denmark, May 2009, pp. 722–726.

[3] J. G. Andrews, H. Claussen, M. Dohler, S. Rangan, and M. C. Reed, "Femtocells: Past, present, and future," in *Proc. IEEE J. Sel. Areas Commun.*, vol. 30, no. 3, pp. 497–508, Apr. 2012.

[4] B. Han *et al.*, "Mobile data offloading through opportunistic communications and social participation," *IEEE Trans. Mobile Comput.*, vol. 11, no. 5, pp. 821–834, May 2012.

[5] "Data Offload–connecting intelligently," White Paper, Juniper Research, Hampshire, U.K., 2013.

[6] M. H. Cheung, R. Southwell, and J. Huang, "Congestion-aware network selection and data offloading," in *Proc. IEEE Annu. Conf. Inf. Sci. Syst. (CISS)*, Princeton, NJ, USA, Mar. 2014, pp. 1–6.

[7] A. Aijaz, H. Aghvami, and M. Amani, "A survey on mobile data offloading: Technical and business perspectives," *IEEE Wireless Commun.*, vol. 20, no. 2, pp. 104–112, Apr. 2013.

[8] F. Rebecchi *et al.*, "Data offloading techniques in cellular networks: A survey," *IEEE Commun. Surveys Tuts.*, vol. 17, no. 2, pp. 580–603, 2nd Quart. 2015.

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

[10] E. Hossain, M. Rasti, H. Tabassum, and A. Abdelnasser, "Evolution toward 5G multi-tier cellular wireless networks: An interference management perspective," *IEEE Wireless Commun.*, vol. 21, no. 3, pp. 118–127, Jun. 2014.

[11] A. Zakrzewska, S. Ruepp, and M. S. Berger, "Towards converged 5G mobile networks-challenges and current trends," in *Proc. ITU Kaleidoscope Acad. Conf.*, St. Petersburg, FL, USA, Jun. 2014, pp. 39–45.

[12] M. N. Tehrani, M. Uysal, and H. Yanikomeroglu, "Device-to-device communication in 5G cellular networks: Challenges, solutions, and future directions," *IEEE Commun. Mag.*, vol. 52, no. 5, pp. 86–92, May 2014.

[13] S. Talwar *et al.*, "Enabling technologies and architectures for 5G wireless," in *Proc. IEEE Microw. Symp. (IMS)*, Tampa, FL, USA, Jun. 2014, pp. 1–4.

[14] O. E. Falowo and H. A. Chan, "RAT selection for multiple calls in heterogeneous wireless networks using modified topsis group decision making technique," in *Proc. IEEE Pers. Indoor Mobile Radio Commun. (PIMRC)*, Toronto, ON, Canada, Sep. 2011, pp. 1371–1375.

[15] K. Yang, I. Gondal, B. Qiu, and L. S. Dooley, "Combined SINR based vertical handoff algorithm for next generation heterogeneous wireless networks," in *Proc. Globecom*, Washington, DC, USA, Nov. 2007, pp. 4483–4487.

[16] H. J. Wang, R. H. Katz, and J. Giese, "Policy-enabled handoffs across heterogeneous wireless networks," in *Proc. IEEE Mobile Comput. Syst. Appl.*, New Orleans, LA, USA, Feb. 1999, pp. 51–60.

[17] Q. Guo, J. Zhu, and X. Xu, "An adaptive multi-criteria vertical handoff decision algorithm for radio heterogeneous network," in *Proc. IEEE Int. Conf. Commun. (ICC)*, vol. 4. Seoul, South Korea, May 2005, pp. 2769–2773.

[18] F. Zhu and J. McNair, "Optimizations for vertical handoff decision algorithms," in *Proc. IEEE WCNC*, vol. 2. Atlanta, GA, USA, Mar. 2004, pp. 867–872.

[19] H. Liu, C. Maciocco, V. Kesavan, and A. L. Y. Low, "Energy efficient network selection and seamless handovers in mixed networks," in *Proc. World Wireless Mobile Multimedia Netw. Workshops (WoWMoM)*, Kos, Greece, Jun. 2009, pp. 1–9.

[20] N. Ristanovic, J. L. Boudec, A. Chaintreau, and V. Erramilli, "Energy efficient offloading of 3G networks," in *Proc. IEEE Mobile Adhoc Sensor Syst. (MASS)*, Valencia, Spain, Oct. 2011, pp. 202–211.

[21] D. S. Deif, H. El-Badawy, and H. El-Hennawy, "Topology based modeling and simulation of UMTS-WLAN wireless heterogeneous network," in *Proc. IEEE Wireless Opt. Commun. Netw. (WOCN)*, Colombo, Sri Lanka, Sep. 2010, pp. 1–5.

[22] 3GPP TS 24.312 v. 11.5.0, "Access network discovery and selection function (ANDSF) management object (MO)," Dec. 2012.

[23] 3GPP TS 24.312 v. 11.5.0, "Architecture enhancements for non-3GPP accesses," Dec. 2012.

[24] D. H. Hagos and R. Kapitza, "Study on performance-centric offload strategies for LTE networks," in *Proc. IEEE Wireless Mobile Netw. Conf. (WMNC)*, Dubai, United Arab Emirates, Apr. 2013, pp. 1–10.

[25] D. S. Kim, Y. Noishiki, Y. Kitatsuji, and H. Yokota, "Efficient ANDSF-assisted Wi-Fi control for mobile data offloading," in *Proc. Int. Wireless Commun. Mobile Comput. Conf. (IWCMC)*, Jul. 2013, pp. 343–348.

[26] J. Huang *et al.*, "A close examination of performance and power characteristics of 4G LTE networks," in *Proc. 10th Int. Conf. Mobile Syst. Appl. Services*, 2012, pp. 225–238.

[27] K. Doppler, C. B. Ribeiro, and J. Kneckt, "On efficient discovery of next generation local area networks," in *Proc. IEEE Wireless Commun. Netw. Conf. (WCNC)*, Cancún, Mexico, Mar. 2011, pp. 269–274.

[28] K. Pahlavan *et al.*, "Handoff in hybrid mobile data networks," *IEEE Pers. Commun.*, vol. 7, no. 2, pp. 34–47, Apr. 2000.

[29] B. D. Higgins *et al.*, "Intentional networking: Opportunistic exploitation of mobile network diversity," in *Proc. MobiCom*, Chicago, IL, USA, Sep. 2010, pp. 73–84.

[30] A. AL Sabbagh, R. Braun, and M. Abolhasan, "Centralized and distributed CRRM in heterogeneous wireless networks," in *Advanced Methods and Applications in Computational Intelligence* (Topics in Intelligent Engineering and Informatics), vol. 6. Cham, Switzerland: Springer Int., Jan. 2014, pp. 299–314.

[31] R. Agusti, J. P. Romero, O. Sallent, and M. Diaz-Guerra, *Radio Resource Management Strategies in UMTS*. Chichester, U.K.: Wiley, 2005.

[32] F. Bari and V. C. M. Leung, "Automated network selection in a heterogeneous wireless network environment," *IEEE Netw.*, vol. 21, no. 1, pp. 34–40, Jan./Feb. 2007.

[33] A. Mihovska *et al.*, "Requirements and algorithms for cooperation of heterogeneous radio access networks," *Wireless Pers. Commun.*, vol. 50, no. 2, pp. 207–245, Jul. 2009.

[34] S. Buljore *et al.*, "IEEE P1900.4 system overview on architecture and enablers for optimised radio and spectrum resource usage," in *Proc. IEEE Symp. New Front. Dyn. Spectr. Access Netw.*, Chicago, IL, USA, Oct. 2008, pp. 1–8.

[35] J. Perez-Romero, O. Sallent, R. Agusti, J. Nasreddine, and M. Muck, "Radio access technology selection enabled by IEEE P1900.4," in *Proc. IST Mobile Wireless Commun. Summit*, Budapest, Hungary, Jul. 2007, pp. 1–5.

[36] A. Tolli, P. Hakalin, and H. Holma, "Performance evaluation of common radio resource management (CRRM)," in *Proc. IEEE Int. Conf. Commun. (ICC)*, vol. 5. New York, NY, USA, 2002, pp. 3429–3433.

[37] D. Suh, H. Ko, and S. Pack, "Efficiency analysis of WiFi offloading techniques," *IEEE Trans. Veh. Technol.*, to be published.

[38] J. Kou, J. Miao, Y. Xiao, Y. Wang, and D. A. Saikrishna, "An offloading algorithm based on channel quality in mobile integration network," in *Proc. ICSP*, Hangzhou, China, Oct. 2014, pp. 1735–1738.

[39] F. Malandrino, C. Casetti, and C.-F. Chiasserini, "LTE offloading: When 3GPP policies are just enough," in *Proc. Wireless On-Demand Netw. Syst. Services (WONS)*, Obergurgl, Austria, Apr. 2014, pp. 1–8.

[40] Y. Choi, H. W. Ji, J.-Y. Park, H.-C. Kim, and J. A. Silvester, "A 3W network strategy for mobile data traffic offloading," *IEEE Commun. Mag.*, vol. 49, no. 10, pp. 118–123, Oct. 2011.

[41] E. Bulut and B. K. Szymanski, "WiFi access point deployment for efficient mobile data offloading," *ACM SIGMOBILE Mobile Comput. Commun. Rev.*, vol. 17, no. 1, pp. 71–78, Jan. 2013.

[42] E. M. R. Oliveira and A. Carneiro, "Routine-based network deployment," in *Proc. INFOCOM Workshops*, Toronto, ON, Canada, May 2014, pp. 183–184.

[43] P. Fuxjäger, H. R. Fischer, I. Gojmerac, and P. Reichl, "Radio resource allocation in urban femto-WiFi convergence scenarios," in *Proc. EURO-NF Conf. Next Gener. Internet (NGI)*, Paris, France, Jun. 2010, pp. 1–8.

[44] S. P. Thiagarajah, A. Ting, D. Chieng, M. Y. Alias, and T. S. Wei, "User data rate enhancement using heterogeneous LTE-802.11n offloading in urban area," in *Proc. ISWTA*, Kuching, Malaysia, Sep. 2013, pp. 11–16.

[45] A. R. Elsherif, W.-P. Chen, A. Ito, and Z. Ding, "Adaptive small cell access of licensed and unlicensed bands," in *Proc. IEEE Int. Conf. Commun. (ICC)*, Budapest, Hungary, Jun. 2013, pp. 6327–6332.

[46] J. Kim, N.-O. Song, B. H. Jung, H. Leem, and D. K. Sung, "Placement of WiFi access points for efficient WiFi offloading in an overlay network," in *Proc. Pers. Indoor Mobile Radio Commun. (PIMRC)*, London, U.K., Sep. 2013, pp. 3066–3070.

[47] B. H. Jung, N.-O. Song, and D. K. Sung, "A network-assisted user-centric WiFi-offloading model for maximizing per-user throughput in a heterogeneous network," *IEEE Trans. Veh. Technol.*, vol. 63, no. 4, pp. 1940–1945, May 2013.

[48] L. G. U. Garcia, I. Rodriguez, D. Catania, and P. Mogensen, "IEEE 802.11 networks: A simple model geared towards offloading studies and considerations on future small cells," in *Proc. Veh. Technol. Conf. (VTC)*, Las Vegas, NV, USA, Sep. 2013, pp. 1–6.

[49] A. Roy and A. Karandikar, "Optimal radio access technology selection policy for LTE-WiFi network," in *Proc. Model. Optim. Mobile Ad Hoc Wireless Netw. (WiOpt)*, Mumbai, India, May 2015, pp. 291–298.

[50] O. Shoukry *et al.*, "Proactive scheduling for content pre-fetching in mobile networks," in *Proc. IEEE Int. Conf. Commun. (ICC)*, Sydney, NSW, Australia, Jun. 2014, pp. 2848–2854.

[51] N. Cheng, N. Lu, N. Zhang, X. S. Shen, and J. W. Mark, "Opportunistic WiFi offloading in vehicular environment: A queueing analysis," in *Proc. IEEE Glob. Commun. Conf. (GLOBECOM)*, Austin, TX, USA, Dec. 2014, pp. 211–216.

[52] A. Balasubramanian, R. Mahajan, and A. Venkataramani, "Augmenting mobile 3G using WiFi," in *Proc. MobiSys*, San Francisco, CA, USA, Jun. 2010, pp. 209–221.

[53] 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. 536–551, Apr. 2013.

[54] E. M. Mohamed, K. Sakaguchi, and S. Sampei, "Delayed offloading zone associations using cloud cooperated heterogeneous networks," in *Proc. IEEE Wireless Commun. Netw. Conf. Workshops (WCNCW)*, New Orleans, LA, USA, Mar. 2015, pp. 374–379.

[55] D. Zhang and C. K. Yeo, "Optimal handing-back point in mobile data offloading," in *Proc. IEEE Veh. Netw. Conf. (VNC)*, Seoul, South Korea, Nov. 2012, pp. 219–225.

[56] International Telecommunications Union, "Guidelines for evaluation of radio-interface technologies for IMT-advanced," Radiocommun. Bureau, Geneva, Switzerland, Tech. Rep. ITU-R M.2135-1, Dec. 2009.

[57] International Telecommunications Union, "Propagation data and prediction methods for the planning of short-range outdoor radiocommunication systems and radio local area networks in the frequency range 300MHz to 100GHz," Radiocommun. Bureau, Geneva, Switzerland, Tech. Rep. ITU-R M.1411-5, Dec. 2009.

[58] L. Hu *et al.*, "Realistic indoor Wi-Fi and femto deployment study as the offloading solution to LTE macro network," in *Proc. IEEE Veh. Technol. Conf. (VTC Fall)*, Quebec City, QC, Canada, Sep. 2012, pp. 1–6.

[59] N. Laoutaris, G. Smaragdakis, P. Rodriguez, and R. Sundaram, "Delay tolerant bulk data transfers on the Internet," in *Proc. 11th Int. Joint Conf. Measur. Model. Comput. Syst. (SIGMETRICS)*, Seattle, WA, USA, Jun. 2009, pp. 229–238.

[60] 3GPP TS 124.312 v. 12.0, "Access network discovery and selection function (ANDSF) management object (MO)," document ETSI TS 124 312, Eur. Telecommun. Standards Inst., Sophia Antipolis, France, 2013.

[61] J. E. Beasley, *Advances in Linear and Integer Programming*. New York, NY, USA: Oxford Univ. Press, 1996.

[62] *IBM ILOG CPLEX Optimizer*. Accessed on Jul. 29, 2015. [Online]. Available: https://www.ibm.com/developerworks/downloads/ws/ilogcplex/

[63] G. Iosifidis, L. Gao, J. Huang, and L. Tassiulas, "An iterative double auction for mobile data offloading," in *Proc. Model. Optim. Mobile Ad Hoc Wireless Netw. (WiOpt)*, Tsukuba, Japan, May 2013, pp. 154–161.

[64] X. Zhuo, W. Gao, G. Cao, and S. Hua, "An incentive framework for cellular traffic offloading," *IEEE Trans. Mobile Comput.*, vol. 13, no. 3, pp. 541–555, Mar. 2014.

[65] X. Zhuo, W. Gao, G. Cao, and Y. Dai, "Win-Coupon: An incentive framework for 3G traffic offloading," in *Proc. ICNP*, Vancouver, BC, Canada, Oct. 2011, pp. 206–215.

[66] 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.

[67] L. Qiu, H. Rui, and A. Whinston, "When cellular capacity meets WiFi hotspots: A smart auction system for mobile data offloading," in *Proc. Hawaii Int. Conf. Syst. Sci. (HICSS)*, Jan. 2015, pp. 4898–4907.

[68] L. Gao, G. Iosifidis, J. Huang, L. Tassiulas, and D. Li, "Bargaining-based mobile data offloading," *IEEE J. Sel. Areas Commun.*, vol. 32, no. 6, pp. 1114–1125, Jun. 2014.

[69] C. Joe-Wong, S. Sen, and S. Ha, "Offering supplementary wireless technologies: Adoption behavior and offloading benefits," in *Proc. IEEE INFOCOM*, Turin, Italy, Apr. 2013, pp. 1061–1069.

[70] L. Gao, G. Iosifidis, J. Huang, and L. Tassiulas, "Economics of mobile data offloading," in *Proc. INFOCOM*, Turin, Italy, Apr. 2013, pp. 3303–3308.

[71] J. Lee, Y. Yi, S. Chong, and Y. Jin, "Economics of WiFi offloading: Trading delay for cellular capacity," *IEEE Trans. Wireless Commun.*, vol. 13, no. 3, pp. 1540–1554, Mar. 2014.

[72] J. Lee, C. Shao, H. Roh, and W. Lee, "Price-based tethering for cooperative networking," in *Proc. Int. Conf. Inf. Netw. (ICOIN)*, Bangkok, Thailand, Jan. 2013, pp. 379–384.

[73] C. Joe-Wong, S. Sen, and S. Ha, "Offering supplementary network technologies: Adoption behavior and offloading benefits," *IEEE/ACM Trans. Netw.*, vol. 23, no. 2, pp. 355–368, Apr. 2015.

[74] K. Poularakis, G. Iosifidis, and L. Tassiulas, "A framework for mobile data offloading to leased cache-endowed small cell networks," in *Proc. IEEE Mobile Ad Hoc Sensor Syst. (MASS)*, Philadelphia, PA, USA, Oct. 2014, pp. 327–335.

[75] Z. Zhou *et al.*, "Data offloading in two-tier networks: A contract design approach," in *Proc. IEEE Glob. Commun. Conf. (GLOBECOM)*, Austin, TX, USA, Dec. 2014, pp. 4531–4536.

[76] X. Kang, Y.-K. Chia, S. Sun, and H. F. Chong, "Mobile data offloading through a third-party WiFi access point: An operator's perspective," *IEEE Trans. Wireless Commun.*, vol. 13, no. 10, pp. 5340–5351, Oct. 2014.

[77] Y. Im *et al.*, "AMUSE: Empowering users for cost-aware offloading with throughput-delay tradeoffs," in *Proc. INFOCOM*, Turin, Italy, Apr. 2013, pp. 435–439.

[78] M. El Chamie, C. Barakat, and G. Neglia, "Geographically fair in-network caching for mobile data offloading," in *Proc. IFIP Netw. Conf. (IFIP Netw.)*, Toulouse, France, May 2015, pp. 1–9.

[79] S. Paris, F. Martignon, I. Filippini, and L. Chen, "A bandwidth trading marketplace for mobile data offloading," in *Proc. INFOCOM*, Turin, Italy, Apr. 2013, pp. 430–434.

[80] T. Han and N. Ansari, "Enabling mobile traffic offloading via energy spectrum trading," *IEEE Trans. Wireless Commun.*, vol. 13, no. 6, pp. 3317–3328, Jun. 2014.

[81] A. Apostolaras, G. Iosifidis, K. Chounos, T. Korakis, and L. Tassiulas, "C2M: Mobile data offloading to mesh networks," in *Proc. IEEE Glob. Commun. Conf. (GLOBECOM)*, Austin, TX, USA, Dec. 2014, pp. 4877–4883.

[82] A. Kliks, N. Dimitriou, A. Zalonis, and O. Holland, "WiFi traffic offloading for energy saving," in *Proc. 20th Int. Conf. Telecommun. (ICT)*, Casablanca, Morocco, May 2013, pp. 1–5.

[83] S. Taleb, M. Dia, J. Farhat, Z. Dawy, and H. Hajj, "On the design of energy-aware 3G/WiFi heterogeneous networks under realistic conditions," in *Proc. Adv. Inf. Netw. Appl. Workshops (WAINA)*, Barcelona, Spain, Mar. 2013, pp. 523–527.

[84] V. A. Siris and M. Anagnostopoulou, "Performance and energy efficiency of mobile data offloading with mobility prediction and prefetching," in *Proc. World Wireless Mobile Multimedia Netw. (WoWMoM)*, Madrid, Spain, Jun. 2013, pp. 1–6.

[85] M. I. Sanchez, C. J. Bernardos, A. De La Oliva, and P. Serrano, "Energy consumption savings with 3G offload," in *Proc. Veh. Technol. Conf. (VTC Fall)*, Las Vegas, NV, USA, Sep. 2013, pp. 1–5.

[86] S. Sharafeddine, K. Jahed, N. Abbas, E. Yaacoub, and Z. Dawy, "Exploiting multiple wireless interfaces in smartphones for traffic offloading," in *Proc. Commun. Netw. (BlackSeaCom)*, Batumi, Georgia, Jul. 2013, pp. 142–146.

[87] M. Segata, B. Bloessl, C. Sommer, and F. Dressler, "Towards energy efficient smart phone applications: Energy models for offloading tasks into the cloud," in *Proc. IEEE Int. Conf. Commun. (ICC)*, Sydney, NSW, Australia, Jun. 2014, pp. 2394–2399.

[88] N. Ristanovic, J.-Y. Le Boudec, A. Chaintreau, and V. Erramilli, "Energy efficient offloading of 3G networks," in *Proc. IEEE 8th Int. Conf. Mobile Ad-Hoc Sensor Syst. (MASS)*, Valencia, Spain, Oct. 2011, pp. 202–211.

[89] A. Y. Ding *et al.*, "Enabling energy-aware collaborative mobile data offloading for smartphones," in *Proc. IEEE Sensor Mesh Ad Hoc Commun. Netw. (SECON)*, New Orleans, LA, USA, Jun. 2013, pp. 487–495.

[90] S. Chen, Z. Yuan, and G.-M. Muntean, "An energy-aware multipath-TCP-based content delivery scheme in heterogeneous wireless networks," in *Proc. IEEE Wireless Commun. Netw. Conf. (WCNC)*, Shanghai, China, Apr. 2013, pp. 1291–1296.

[91] V. Miliotis, L. Alonso, and C. Verikoukis, "Energy efficient proportionally fair uplink offloading for IP flow mobility," in *Proc. IEEE 19th Int. Workshop Comput. Aided Model. Design Commun. Links Netw. (CAMAD)*, Athens, Greece, Dec. 2014, pp. 6–10.

[92] F.-T. Hsu and H.-J. Su, "When does the AP deployment incentivize a user to offload cellular data: An energy efficiency viewpoint," in *Proc. Commun. Control Signal Process. (ISCCSP)*, Athens, Greece, May 2014, pp. 210–213.

[93] B. H. Jung, N.-O. Song, and D. K. Sung, "An energy-efficient WiFi offloading model in a heterogeneous network," in *Proc. Green Commun. (OnlineGreencomm)*, Tucson, AZ, USA, Nov. 2014, pp. 1–5.

[94] U. Sethakaset, Y.-K. Chia, and S. Sun, "Energy efficient WiFi offloading for cellular uplink transmissions," in *Proc. Veh. Technol. Conf. (VTC Spring)*, Seoul, South Korea, May 2014, pp. 1–5.

[95] N. Abbas, Z. Dawy, H. Hajj, and S. Sharafeddine, "Energy-throughput tradeoffs in cellular/WiFi heterogeneous networks with traffic splitting," in *Proc. IEEE Wireless Commun. Netw. Conf. (WCNC)*, Istanbul, Turkey, Apr. 2014, pp. 2294–2299.

[96] M. Kuhnert and C. Wietfeld, "Performance evaluation of an advanced energy-aware client-based handover solution in heterogeneous LTE and WiFi networks," in *Proc. Veh. Technol. Conf. (VTC Spring)*, Seoul, South Korea, May 2014, pp. 1–5.

[97] V. Miliotis, L. Alonso, and C. Verikoukis, "Offloading with IFOM: The uplink case," in *Proc. Glob. Commun. Conf. (GLOBECOM)*, Austin, TX, USA, Dec. 2014, pp. 2661–2666.

[98] M. Simsek, M. Bennis, M. Debbah, and A. Czylwik, "Rethinking offload: How to intelligently combine WiFi and small cells?" in *Proc. IEEE Int. Conf. Commun. (ICC)*, Budapest, Hungary, Jun. 2013, pp. 5204–5208.

[99] M. Hu, J. Zhang, and J. Sadowsky, "Traffic aided opportunistic scheduling for wireless networks: Algorithms and performance bounds," *Comput. Netw.*, vol. 46, no. 4, pp. 505–518, Nov. 2004.

[100] C. K. Ho, D. Yuan, and S. Sun, "Data offloading in load coupled networks: A utility maximization framework," *IEEE Trans. Wireless Commun.*, vol. 13, no. 4, pp. 1921–1931, Apr. 2014.

[101] C. K. Ho, D. Yuan, and S. Sun, "Data offloading in load coupled networks: Solution characterization and convexity analysis," in *Proc. IEEE Int. Conf. Commun. Workshops (ICC)*, Budapest, Hungary, Jun. 2013, pp. 1161–1165.

[102] S. Singh, H. S. Dhillon, and J. G. Andrews, "Offloading in heterogeneous networks: Modeling, analysis, and design insights," *IEEE Trans. Wireless Commun.*, vol. 12, no. 5, pp. 2484–2497, Apr. 2013.

[103] O. Galinina, A. Pyattaev, S. Andreev, M. Dohler, and Y. Koucheryavy, "5G multi-RAT LTE-WiFi ultra-dense small cells: Performance dynamics, architecture, and trends," *IEEE J. Sel. Areas Commun.*, vol. 33, no. 6, pp. 1224–1240, Jun. 2015.

[104] D. P. Bertsekas, D. A. Castanon, and H. Tsaknakis, "Reverse auction and the solution of inequality constrained assignment problems," *SIAM J. Optim.*, vol. 3, no. 2, pp. 268–299, Mar. 1992.

[105] L. Song, D. Kotz, R. Jain, and X. He, "Evaluating next-cell predictors with extensive Wi-Fi mobility data," *IEEE Trans. Mobile Comput.*, vol. 5, no. 12, pp. 1633–1649, Dec. 2006.

[106] (Feb. 2012). *Integrated Femto-WiFi Networks*. [Online]. Available: http://www.smallcellforum.org

[107] S. A. AlQahtani, "Analysis of resource splitting scheme with cognitive based admission control for femto-WiFi wireless networks," *Wireless Netw.*, vol. 20, no. 8, pp. 2307–2317, Jun. 2014.

[108] S.-P. Yeh, S. Talwar, G. Wu, N. Himayat, and K. Johnsson, "Capacity and coverage enhancement in heterogeneous networks," *IEEE Wireless Commun.*, vol. 18, no. 3, pp. 32–38, Jun. 2011.

[109] A. Kliks, N. Dimitriou, A. Zalonis, and O. Holland, "WiFi traffic offloading for energy saving," in *Proc. IEEE Int. Conf. Telecommun. (ICT)*, Casablanca, Morocco, May 2013, pp. 1–5.

[110] *IP Flow Mobility and Seamless Wireless Local Area Network (WLAN) Offload; Stage 2 (v11.0.0)*, document 3GPP TS 23.261, 3rd Gener. Partnership Project, Eur. Telecommun. Stand. Inst., Sophia Antipolis, France, Sep. 2012.

[111] C. B. Sankaran, "Data Offloading techniques in 3GPP Rel-10 networks: A tutorial," *IEEE Commun. Mag.*, vol. 50, no. 6, pp. 46–53, Jun. 2012.

[112] G. Bianchi, "Performance analysis of the IEEE 802.11 distributed coordination function," *IEEE J. Sel. Topics Signal Process.*, vol. 18, no. 3, pp. 535–547, Mar. 2000.

[113] S. Gundavelli, K. Leung, V. Devarapalli, K. Chowdhury, and B. Patil, *Proxy Mobile IPv6*, document RFC 5213, Internet Eng. Task Force, Fremont, CA, USA, Aug. 2008.

[114] R. S. Sutton and A. G. Barto, "Reinforcement Learning: An Introduction," *IEEE Trans. Neural Netw.*, vol. 9, no. 5, p. 1054, Sep. 1998.

[115] L. E. Schrage and L. W. Miller, "The queue M/G/1 with the shortest remaining processing time discipline," *Oper. Res.*, vol. 14, no. 4, pp. 670–684, Aug. 1966.

[116] S. Borst and P. Whiting, "Dynamic rate control algorithms for HDR throughput optimization," in *Proc. IEEE INFOCOM*, Anchorage, AK, USA, Apr. 2001, pp. 976–985.

[117] X. Liu, E. K. P. Chong, and N. B. Shroff, "Opportunistic transmission scheduling with resource-sharing constraints in wireless networks," *IEEE J. Sel. Area Commun.*, vol. 19, no. 10, pp. 2053–2064, Oct. 2001.

[118] I. Siomina and D. Yuan, "Analysis of cell load coupling for LTE network planning and optimization," *IEEE Trans. Wireless Commun.*, vol. 11, no. 6, pp. 2287–2297, Jun. 2012.

[119] R. S. Kumar and S. Ramachandram, "Load balancing in genetic zone routing protocol for MANETs," *Int. J. Comput. Inf. Eng.*, vol. 3, no. 4, pp. 261–266, 2009.

[120] S. Budiyanto, M. Asvial, and D. Gunawan, "Implementation of genetic zone routing protocol (GZRP) in 3G-WiFi offload multi base station," in *Proc. IEEE TENCON*, Xi'an, China, Oct. 2013, pp. 1–6.

[121] R. C. Chalmers and K. C. Almeroth, "A mobility gateway for small-device networks," in *Proc. IEEE Int. Conf. Pervasive Comput. Commun.*, Orlando, FL, USA, Mar. 2004, pp. 209–218.

[122] N. Thompson, G. He, and H. Luo, "Flow scheduling for end-host multihoming," in *Proc. INFOCOM*, Barcelona, Spain, Apr. 2006, pp. 1–12.

[123] X. Wu, M. C. Chan, and A. L. Ananda, "TCP handoff: A practical TCP enhancement for heterogeneous mobile environments," in *Proc. IEEE Int. Conf. Commun. (ICC)*, Glasgow, U.K., Jun. 2007, pp. 6043–6048.

[124] S.-E. Kim and J. A. Copeland, "TCP for seamless vertical handoff in hybrid mobile data networks," in *Proc. IEEE Glob. Telecommun. Conf. (GLOBECOM)*, vol. 2. San Francisco, CA, USA, Dec. 2003, pp. 661–665.

[125] S. Nirjon, A. Nicoara, C.-H. Hsu, J. Singh, and J. Stankovic, "MultiNets: Policy oriented real-time switching of wireless interfaces on mobile devices," in *Proc. IEEE 16th Real Time Embedded Technol. App. Symp.*, Beijing, China, Apr. 2012, pp. 251–260.

[126] Q. Li, Q. Han, and L. Sun, "Context-aware handoff on smartphones," in *Proc. IEEE Mobile Ad-Hoc Sensor Syst. (MASS)*, Oct. 2013, pp. 470–478.

[127] J. Howe, "The rise of crowdsourcing," *Wired Mag.*, vol. 14, no. 6, pp. 1–4, 2006.

[128] C. M. Bishop, *Pattern Recognition and Machine Learning* (Information Science and Statistics). New York, NY, USA: Springer-Verlag, 2006.

[129] S. Andreev, A. Pyattaev, K. Johnsson, O. Galinina, and Y. Koucheryavy, "Cellular traffic offloading onto network-assisted device-to-device connections," *IEEE Commun. Mag.*, vol. 52, no. 4, pp. 20–31, Apr. 2014.

[130] L. Lei, Z. Zhong, C. Lin, and X. Shen, "Operator controlled device-to-device communications in LTE-advanced networks," *IEEE Wireless Commun.*, vol. 19, no. 3, pp. 96–104, Jun. 2012.

[131] T. Koskela, S. Hakola, T. Chen, and J. Lehtomaki, "Clustering concept using device-to-device communication in cellular system," in *Proc. IEEE Wireless Commun. Netw. Conf.*, Sydney, NSW, Australia, Apr. 2010, pp. 1–6.

[132] H. Han, H. Wang, and X. Lin, "A low-cost link-selection strategy for cellular controlled short-range communications," in *Proc. 12th IEEE Int. Conf. Commun. Technol. (ICCT)*, Nanjing, China, Nov. 2010, pp. 1332–1335.

[133] J. Gu, S. J. Bae, B.-G. Choi, and M. Y. Chung, "Dynamic power control mechanism for interference coordination of device-to-device communication in cellular networks," in *Proc. 3rd Int. Conf. Ubiquitous Future Netw. (ICUFN)*, Dalian, China, Jun. 2011, pp. 71–75.

[134] A. Pyattaev, K. Johnsson, S. Andreev, and Y. Koucheryavy, "3GPP LTE traffic offloading onto WiFi direct," in *Proc. IEEE Wireless Commun. Netw. Conf. Workshops (WCNCW)*, Shanghai, China, Apr. 2013, pp. 135–140.

[135] S. Andreev, O. Galinina, A. Pyattaev, K. Johnsson, and Y. Koucheryavy, "Analyzing assisted offloading of cellular user sessions onto D2D links in unlicensed bands," *IEEE J. Sel. Areas Commun.*, vol. 33, no. 1, pp. 67–80, Jan. 2015.

[136] X. Wang, M. Chen, T. Kwon, L. Jin, and V. C. M. Leung, "Mobile traffic offloading by exploiting social network services and leveraging opportunistic device-to-device sharing," *IEEE Wireless Commun.*, vol. 21, no. 3, pp. 28–36, Jun. 2014.

[137] J. Qiao *et al.*, "Enabling device-to-device communications in millimeter-wave 5G Cellular networks," *IEEE Commun. Mag.*, vol. 53, no. 1, pp. 209–215, Jan. 2015.

[138] D. Karvounas, A. Georgakopoulos, K. Tsagkaris, V. Stavroulaki, and P. Demestichas, "Smart management of D2D constructs: An experiment-based approach," *IEEE Commun. Mag.*, vol. 52, no. 4, pp. 82–89, Apr. 2014.

[139] X. Lu, H. Pan, and P. Lio, "Offloading mobile data from cellular networks through peer-to-peer WiFi communication: A subscribe-and-send architecture," *China Commun.*, vol. 10, no. 6, pp. 35–46, Jun. 2013.

[140] J. Burgess, B. Gallagher, D. Jensen, and B. N. Levine, "MaxProp: Routing for vehicle-based disruption-tolerant networks," in *Proc. 25th IEEE Int. Conf. Comput. Commun. (INFOCOM)*, Barcelona, Spain, Apr. 2006, pp. 1–11.

[141] H. Pan, J. Crowcroft, and E. Yoneki, "Social-based forwarding in delay tolerant networks," in *Proc. 9th ACM Int. Symp. Mobile Ad Hoc Netw. Comput. (MobiHoca)*, Hong Kong, May 2008, pp. 241–250.

[142] P. Kolios, C. Panayiotou, and G. Ellinas, "ExTraCT: Expediting offloading transfers through intervehicle communication transmissions," *IEEE Trans. Intell. Transp. Syst.*, vol. 16, no. 3, pp. 1238–1248, Jun. 2014.

[143] F. H. P. Fitzek, M. Katz, and Q. Zhang, "Cellular controlled short-range communication for cooperative P2P networking," *Wireless Personal Commun.*, vol. 48, no. 1, pp. 141–155, Jan. 2009.

[144] L. Popova, T. Herpel, W. Gerstacker, and W. Koch, "Cooperative mobile-to-mobile file dissemination in cellular networks within a unified radio interface," *Comput. Netw.*, vol. 52, no. 6, pp. 1153–1165, Apr. 2008.

[145] L. Al-Kanj, H. V. Poor, and Z. Dawy, "Optimal cellular offloading via device-to-device communication networks with fairness constraints," *IEEE Trans. Wireless Commun.*, vol. 13, no. 8, pp. 4628–4643, Aug. 2014.

[146] S.-H. Kim, C.-K. Lee, and S.-G. Kang, "Reliable data delivery for relay-based overlay multicast," in *Proc. ICACT*, vol. 1. Feb. 2009, pp. 782–785.

[147] J. Park, J. M. Lee, and S.-G. Kang, "Design and implementation of overlay multicast protocol for many-to-many multicast services," in *Proc. ICACT*, vol. 3. Gangwon-do, South Korea, Feb. 2007, pp. 2144–2147.

[148] *Standard for the Functional Architecture of Next Generation Service Overlay Networks*, IEEE Standard 1903-2011, May 2011.

[149] S.-I. Lee and S.-G. Kang, "NGSON: Features, state of the art, and realization," *IEEE Commun. Mag.*, vol. 50, no. 1, pp. 54–61, Jan. 2012.

[150] J. R. Iyengar, P. D. Amer, and R. Stewart, "Concurrent multipath transfer using SCTP multihoming over independent end-to-end paths," *IEEE/ACM Trans. Netw.*, vol. 14, no. 5, pp. 951–964, Oct. 2006.

[151] S. J. Koh, M. J. Chang, and M. Lee, "mSCTP for soft handover in transport layer," *IEEE Commun. Lett.*, vol. 8, no. 3, pp. 189–191, Mar. 2004.

[152] H. Han, S. Shakkottai, C. V. Hollot, R. Srikant, and D. Towsley, "Multi-path TCP: A joint congestion control and routing scheme to exploit path diversity in the Internet," *IEEE/ACM Trans. Netw.*, vol. 14, no. 6, pp. 1260–1271, Dec. 2006.

[153] T. You, C. Park, H. Jung, T. T. Taekyoung, and Y. Choi, "Multipath transmission architecture for heterogeneous wireless networks," in *Proc. IEEE ICT Convergence*, Seoul, South Korea, Sep. 2011, pp. 26–31.

[154] C. Raiciu, D. Wischik, and M. Handley, "Practical congestion control for multipath transport protocols," Dept. Comput. Sci., Univ. College London, London, U.K., Tech. Rep. 6824, 2011.

[155] M. A. P. Gonzalez, T. Higashino, and M. Okada, "Radio access considerations for data offloading with multipath TCP in cellular/WiFi networks," in *Proc. Inf. Netw. (ICOIN)*, Bangkok, Thailand, Jan. 2013, pp. 680–685.

[156] IETF. (Jul. 2015). *Draft-IETF-MPTCP-Experience-02*. [Online]. Available: http://datatracker.ietf.org/wg/mptcp/documents/

[157] 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. (MASS)*, Oct. 2011, pp. 192–201.

[158] T. Han and N. Ansari, "Opportunistic content pushing via WiFi hotspots," in *Proc. IEEE Int. Conf. Netw. Infrastruct. Digital Content (IC-NIDC)*, Beijing, China, Sep. 2012, pp. 680–684.

[159] B. Han *et al.*, "Cellular traffic offloading through opportunistic communications: A case study," in *Proc. IEEE ACM CHANTS*, Chicago, IL, USA, Sep. 2010, pp. 821–834.

[160] P. Baier, F. Dürr, and K. Rothermel, "TOMP: Opportunistic traffic offloading using movement predictions," in *Proc. IEEE Local Comput. Netw. (LCN)*, Clearwater, FL, USA, Oct. 2012, pp. 50–58.

[161] K. Thilakarathna, A. Seneviratne, A. C. Viana, and H. Petander, "User generated content dissemination in mobile social networks through infrastructure supported content replication," *Pervasive Mobile Comput. J.*, vol. 11, pp. 132–147, Apr. 2014.

[162] K. Thilakarathna, A. C. Viana, A. Seneviratne, and H. Petander, "Mobile social networking through friend-to-friend opportunistic content dissemination," in *Proc. ACM Mobihoc*, Bangalore, India, Aug. 2013, pp. 263–266.

[163] M. V. Barbera, A. C. Viana, M. D. de Amorim, and J. Stefa, "Data offloading in social mobile networks through VIP delegation," *Ad Hoc Netw. J.*, vol. 19, pp. 92–110, Aug. 2014.

[164] K. K. Rachuria, C. Efstratioua, I. Leontiadisa, C. Mascoloa, and P. J. Rentfrowc, "Smartphone sensing offloading for efficiently supporting social sensing applications," *Pervasive Mobile Comput.*, vol. 10, pp. 3–21, Feb. 2014.

[165] J. Ott and D. Kutscher, "The 'drive-thru' architecture: WLAN-based Internet access on the road," in *Proc. IEEE Veh. Technol. Conf. (VTC Spring)*, Milan, Italy, May 2014, pp. 2615–2622.

[166] J. Ott and D. Kutscher, "A disconnection-tolerant transport for drive-thru Internet environments," in *Proc. Annu. Joint Conf. IEEE Comput. Commun. Soc.*, vol. 3. Miami, FL, USA, Mar. 2005, pp. 1849–1862.

[167] J. Eriksson, H. Balakrishnan, and S. Madden, "Cabernet: Vehicular content delivery using WiFi," in *Proc. ACM MobiCom*, San Francisco, CA, USA, Sep. 2008, pp. 199–210.

[168] T. H. Luan, X. Ling, and X. Shen, "MAC in motion: Impact of mobility on the MAC of drive-thru Internet," *IEEE Trans. Mob. Comput.*, vol. 11, no. 2, pp. 305–319, Feb. 2012.

[169] D. Hadaller, S. Keshav, T. Brecht, and S. Agarwal, “Vehicular opportunistic communication under the microscope,” in *Proc. ACM MobiSys*, San Juan, PR, USA, Jun. 2007, pp. 206–219.

[170] A. Giannoulis, M. Fiore, and E. W. Knightly, “Supporting vehicular mobility in urban multi-hop wireless networks,” in *Proc. 6th Int. Conf. Mobile Syst. Appl. Services*, Breckenridge, CO, USA, Jun. 2008, pp. 54–66.

[171] A. Balasubramanian, R. Mahajan, A. Venkataramani, B. N. Levine, and J. Zahorjan, “Interactive WiFi connectivity for moving vehicles,” in *Proc. ACM SIGCOMM Conf. Data Commun.*, Aug. 2008, pp. 427–438.

[172] R. Droms *et al.*, *Dynamic Host Configuration Protocol for IPv6 (DHCPv6)*, document RFC 3315, Internet Eng. Task Force, Fremont, CA, USA, Jul. 2003.

[173] P. Deshpande, A. Kashyap, C. Sung, and S. R. Das, “Predictive methods for improved vehicular WiFi Access,” in *Proc. 7th Int. Conf. Mobile Syst. Appl. Services*, Wrocław, Poland, Jun. 2009, pp. 263–276.

[174] N. Imai, H. Morikawa, and T. Aoyama, “Prefetching architecture for hot-spotted networks,” in *Proc. IEEE Int. Conf. Commun. (ICC)*, vol. 7. Helsinki, Finland, Jun. 2001, pp. 2006–2010.

[175] H. Hartenstein and L. P. Laberteaux, “A tutorial survey on vehicular ad hoc networks,” *IEEE Commun. Mag.*, vol. 46, no. 6, pp. 164–171, Jun. 2008.

[176] M. Gramaglia, C. J. Bernardos, and M. Calderon, “Seamless Internet 3G and opportunistic WLAN vehicular connectivity,” *EURASIP J. Wireless Commun. Netw.*, vol. 2011, no. 183, pp. 1–20, Nov. 2011.

[177] M. Gramaglia, M. Calderon, and C. J. Bernardos, “TREBOL: Tree-based routing and address autoconfiguration for vehicle-to-Internet communications,” in *Proc. IEEE Veh. Technol. Conf. (VTC Spring)*, Budapest, Hungary, May 2011, pp. 1–5.

[178] K. Taniuchi *et al.*, “IEEE 802.21: Media independent handover: Features, applicability, and realization,” *IEEE Commun. Mag.*, vol. 47, no. 1, pp. 112–120, Jan. 2009.

[179] E. Bulut and B. K. Szymanski, “WiFi access point deployment for efficient mobile data offloading,” *ACM SIGMOBILE Mobile Comput. Commun. Rev.*, vol. 17, no. 1, pp. 71–78, Jan. 2013.

[180] B. Zhou, H. Hu, S.-Q. Huang, and H.-H. Chen, “Intracluster device-to-device relay algorithm with optimal resource utilization,” *IEEE Trans. Veh. Technol.*, vol. 62, no. 5, pp. 2315–2326, Jun. 2013.

[181] L. Militano, M. Condoluci, G. Araniti, A. Molinaro, and A. Iera, “When D2D communication improves group oriented services in beyond 4G networks,” *Wireless Netw.*, vol. 21, no. 4, pp. 1363–1377, May 2015.

[182] S. Andreev *et al.*, “Network-assisted device-to-device connectivity: Contemporary vision and open challenges,” in *Proc. 21st Eur. Wireless Conf.*, Budapest, Hungary, May 2015, pp. 1–8.

[183] J. Callaham. (Jul. 2015). *All Nexus 6s on Project Fi Are Due to Receive Android 5.1.1 ‘Over the Next Few Days’*. [Online]. Available: http://www.androidcentral.com/all-project-fi-phones-are-due-receive-android-511-over-next-few-days

[184] L. Ulanoff. (Apr. 2015). *Google Project Fi Is Full of Promise and Questions*. [Online]. Available: http://mashable.com/2015/04/22/google-project-fi-analysis/

[185] *Wi-Fi Certified Passpoint Architecture for Public Access*, Aruba Netw., Sunnyvale, CA, USA, 2012.

[186] S. Hoteit *et al.*, “Mobile data traffic offloading over passpoint hotspots,” *Comput. Netw.*, vol. 84, no. 19, pp. 76–93, Jun. 2015.

[187] A. I. Sulyman *et al.*, “Radio propagation path loss models for 5G cellular networks in the 28 GHz and 38 GHz millimeter-wave bands,” *IEEE Commun. Mag.*, vol. 52, no. 9, pp. 78–86, Sep. 2014.

[188] R. Zhang *et al.*, “LTE-unlicensed: The future of spectrum aggregation for cellular networks,” *IEEE Wireless Commun.*, vol. 22, no. 3, pp. 150–159, Jun. 2015.

[189] H. Hu and Q. Zhu, “Dynamic spectrum access in underlay cognitive radio system with SINR constraints,” in *Proc. Wireless Commun. Network. Mobile Comput.*, Beijing, China, Sep. 2009, pp. 1–4.

[190] C. Ghosh, S. Roy, and M. B. Rao, “Modeling and validation of channel idleness and spectrum availability for cognitive networks,” *IEEE J. Sel. Areas Commun.*, vol. 30, no. 10, pp. 2029–2039, Nov. 2012.

[191] M. Nabil, W. El-Sayed, and M. ElNainay, “A cooperative spectrum sensing scheme based on task assignment algorithm for cognitive radio networks,” in *Proc. IWCMC*, Nicosia, Cyprus, Aug. 2014, pp. 151–156.

[192] C. Lacatus, D. Akopian, P. Yaddanapudi, and M. Shadaram, “Flexible spectrum and power allocation for OFDM unlicensed wireless systems,” *IEEE Syst. J.*, vol. 3, no. 2, pp. 254–264, Jun. 2009.

[193] M. R. Dzulkifli, M. R. Kamarudin, and T. A. Rahman, “Spectrum allocation using genetic algorithm in cognitive radio networks,” in *Proc. IEEERFM*, Dec. 2011, pp. 111–114.

[194] P. Cui, M. Tonnemacher, D. Rajan, and J. Camp, “WhiteCell: Energy-efficient use of unlicensed frequency bands for cellular offloading,” in *Proc. IEEE DySPAN*, Stockholm, Sweden, Sep. 2015, pp. 188–199.

[195] C. W. Kim, J. Ryoo, and M. M. Buddhikot, “Design and implementation of an end-to-end architecture for 3.5 GHz shared spectrum,” in *Proc. IEEE DySPAN*, Stockholm, Sweden, Sep. 2015, pp. 23–34.

[196] (Apr. 2016). *METIS (Mobile and Wireless Communications Enablers for the Twenty-Twenty Information Society)*. [Online]. Available: https://www.metis2020.com

[197] (Apr. 2016). *PHYDYAS (Physical Layer For Dynamic Spectrum Access and Cognitive Radio)*. [Online]. Available: http://www.ict-phvdvas.org/

[198] *EMPhAtiC (Enhanced Multicarrier Techniques for Professional Ad-Hoc and Cell-Based Communications)*. [Online]. Available: http://www.ict-emphatic.eu/

[199] *5GNOW (5th Generation Non-Orthogonal Waveforms for Asynchronous Signalling)*. [Online]. Available: http://www.5gnow.eu/

[200] H. Shimodaira *et al.*, “Cell association method for multiband heterogeneous networks,” in *Proc. Workshop Wireless Distrib. Netw. PMRC*, Washington, DC, USA, Sep. 2014, pp. 2209–2213.

Yejun He (SM’09) received the B.S. degree from the Huazhong University of Science and Technology (HUST), Wuhan, China, in 1994, the M.S. degree from the Wuhan University of Technology, Wuhan, in 2002, and the Ph.D. degree from HUST, in 2005. From 2005 to 2006, he was a Research Associate with the Department of Electronic and Information Engineering, Hong Kong Polytechnic University, Hong Kong. From 2006 to 2007, he was a Research Associate with the Department of Electronic Engineering, Faculty of Engineering, Chinese University of Hong Kong, Hong Kong. In 2012, he was a Visiting Professor with the Department of Electrical and Computer Engineering, University of Waterloo, Waterloo, ON, Canada. From 2013 to 2015, he was an Advanced Visiting Scholar (Visiting Professor) with the School of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, GA, USA. Since 2011, he has been a Professor with the College of Information Engineering, Shenzhen University, Shenzhen, China. He is currently the Director of the Shenzhen Key Laboratory of Antennas and Propagation, Shenzhen. He has authored or co-authored about 100 research papers, and books (chapters), and holds 13 patents. His research interests include channel coding and modulation, 4G/5G wireless mobile communication, space-time processing, antennas, and RF.

Dr. He has been an Associate Editor of *Security and Communication Networks*, since 2012. He was the TPC Co-Chair of WOCC 2015. He has served as a Reviewer for various journals such as the IEEE TRANSACTIONS ON VEHICULAR TECHNOLOGY, the IEEE TRANSACTIONS ON COMMUNICATIONS, the IEEE TRANSACTIONS ON INDUSTRIAL ELECTRONICS, the IEEE WIRELESS COMMUNICATIONS, the IEEE COMMUNICATIONS LETTERS, the *International Journal of Communication Systems*, *Wireless Communications and Mobile Computing*, and *Wireless Personal Communications*. He has also served as a Technical Program Committee Member or a Session Chair for various conferences, including the IEEE Global Telecommunications Conference (GLOBECOM), the IEEE International Conference on Communications (ICC), the IEEE Wireless Communication Networking Conference (WCNC), and the IEEE Vehicular Technology Conference (VTC). He served as an Organizing Committee Vice Chair of the International Conference on Communications and Mobile Computing (CMC 2010) and an Editor of CMC2010 Proceedings. He acted as the Publicity Chair of several international conferences such as the IEEE PIMRC 2012. He is the Principal Investigator for over 20 current or finished research projects, including the NSFC of China, the Integration Project of Production Teaching and Research by Guangdong Province and Ministry of Education as well as the Science and Technology Program of Shenzhen City. He is a Senior Member of the China Institute of Communications and the China Institute of Electronics.

Man Chen (S'13) is currently pursing the M.S. degree in information and communication engineering with Shenzhen University, Shenzhen, China. He has been a Vice Chairman of the IEEE Shenzhen University Student Branch in China Section, since 2014. His research interests include wireless communication systems with a current focus on WiFi offloading strategies, delay-tolerant networking, context-aware offloading, and mobile ad hoc networks, especially vehicular networks.

Baohong Ge is currently pursing the M.S. degree in information and communication engineering with Shenzhen University, Shenzhen, China. Her research interests include wireless communication systems and ad hoc networks.

Mohsen Guizani (S'85–M'89–SM'99–F'09) received the B.S. (with distinction) and M.S. degrees in electrical engineering and the M.S. and Ph.D. degrees in computer engineering from Syracuse University, Syracuse, NY, USA, in 1984, 1986, 1987, and 1990, respectively. He is currently a Professor and the ECE Department Chair with the University of Idaho, ID, USA. He served as an Associate Vice President of Graduate Studies and Research, Qatar University, the Chair of the Computer Science Department, Western Michigan University, the Chair of the Computer Science Department, University of West Florida. He also served in academic positions with the University of Missouri–Kansas City, the University of Colorado-Boulder, Syracuse University, and Kuwait University. He has authored 9 books and over 400 publications in refereed journals and conferences. His research interests include wireless communications and mobile computing, computer networks, mobile cloud computing, security, and smart grid. He currently serves on the editorial boards of several international technical journals and is the Founder and Editor-in-Chief of *Wireless Communications and Mobile Computing* (Wiley). He was a Guest Editor of a number of special issues in IEEE journals and magazines. He also served as a member, Chair, and General Chair of a number of international conferences. He was selected as the Best Teaching Assistant for two consecutive years at Syracuse University. He was the Chair of the IEEE Communications Society Wireless Technical Committee and the Chair of the TAOS Technical Committee. He served as the IEEE Computer Society Distinguished Speaker from 2003 to 2005.