Data Offloading Techniques in Cellular Networks: A Survey

Filippo Rebecchi, Marcelo Dias de Amorim, Vania Conan, Andrea Passarella, Raffaele Bruno, Marco Conti

Data Offloading Techniques in Cellular Networks: A Survey

Filippo Rebecchi, Marcelo Dias de Amorim, Vania Conan, Andrea Passarella, Raffaele Bruno, and Marco Conti

***Abstract*—One of the most engaging challenges for mobile operators today is how to manage the exponential data traffic increase. Mobile data offloading stands out as a promising and low cost solution to reduce the burden on the cellular network. To make this possible, we need a new hybrid network paradigm that leverages the existence of multiple alternative communication channels. This entails significant modifications in the way data is handled, affecting also the behavior of network protocols. In this paper, we present a comprehensive survey of data offloading techniques in cellular networks and extract the main requirements needed to integrate data offloading capabilities into today's mobile networks. We classify existing strategies into two main categories, according to their requirements in terms of content delivery guarantees: delayed and non-delayed offloading. We overview the technical aspects and discuss the state of the art in each category. Finally, we describe in detail the novel functionalities needed to implement mobile data offloading in the access network, as well as current and future research challenges in the field, with an eye toward the design of hybrid architectures.**

***Index Terms*—Mobile data offloading, hybrid networks, WiFi, delay-tolerant networks, cellular networks.**

I. Introduction

GLOBAL mobile traffic will boom in the years to come, thanks to the increasing popularity of smart mobile devices and the introduction of affordable data plans by cellular operators. Data hungry mobile applications, such as audio and video streaming, social sharing, or cloud-based services, are more and more popular among users. Recently, analysts from Cisco warned that global mobile data traffic is expected to grow 18-fold between 2011 and 2018, three times faster than the overall fixed IP traffic in the same period [1]. It is also anticipated that 66.5% of this traffic will be video related (with or without real-time requirements) by 2017. As today's most common data access method on the move, cellular networks are under pressure trying to cope with this unprecedented data overload. Accommodating this growth requires major investments both in the radio access network (RAN) and the core infrastructures. From a purely economic perspective, upgrading the RAN is very expensive, since this approach requires more infrastructure equipment and thus more investment.

Scarce licensed spectrum hinders the RAN enhancements. Regulations allow mobile operators to use only a small portion of the overall radio spectrum, which is also extremely expensive. Users must share the same limited wireless resources. Adding traffic beyond a certain limit mines the performance and the quality of service (QoS) perceived by the users. During peak times in crowded metropolitan environments, users already experience long latencies, low throughput, and network outages due to congestion and overload at RAN level [2]. Unfortunately, this trend can only exacerbate in future due to the predicted mobile data explosion. The problem concerns primarily network operators because they have to trade-off customer satisfaction with business profitability, given the trend toward nearly flat rate business models. In other words, the exponential increase in traffic flowing in their RAN does not generate enough additional revenues to be allocated into further RAN upgrades. This creates what Mölleryd et al. call the *revenue gap* [3].

The above-mentioned circumstances fostered the interest in alternative methods to mitigate the pressure on the cellular network. As a first option, mobile operators solved this contingency by throttling connection speed and capping data usage [4]. However, these practices negatively affect the customer satisfaction. For this reason, alternative approaches emerged. In this survey, we turn our attention to one of these solutions, recently attracting increasing interest by the research community: *mobile data offloading*. An intuitive approach is to leverage the unused bandwidth across different wireless technologies. We consider mobile data offloading as the use of *a complementary wireless technology to transfer data originally targeted to flow through the cellular network*, in order to improve some key performance indicators.

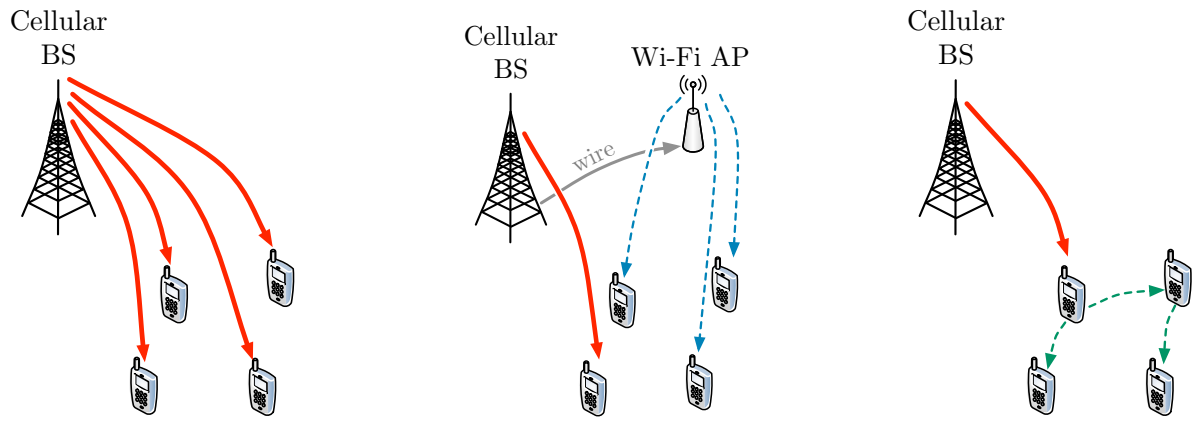
Although offloading may apply to any network, current academic and industrial research mostly concerns with offloading data from cellular networks. Those are the type of networks that would benefit most from this technique. Note that, for the sake of tractability, we limit the scope of this survey to solutions in which mobile terminals are explicitly used as part of the offloading scheme, either through using multiple wireless interfaces, or through using non-conventional cellular

F. Rebecchi is with LIP6 – UPMC Sorbonne Universits and Thales Communications & Security, France (filippo.rebecchi@lip6.fr).

M. Dias de Amorim is with CNRS/UPMC Sorbonne Universits, 4 Pl. Jussieu, 75005 Paris, France (marcelo.amorim@lip6.fr).

V. Conan is with Thales Communications & Security, 4 Av. des Louvresses, 92230 Gennevilliers, France (vania.conan@thalesgroup.com).

A. Passarella, R. Bruno, and M. Conti are with IIT–CNR, Via Giuseppe Moruzzi, 1, 56124 Pisa, Italy ({a.passarella, raffaele.bruno, marco.conti}@iit.cnr.it).



(a) Infrastructure only. (b) AP-based offloading. (c) Terminal-to-terminal offloading.

Fig. 1. The two major approaches to cellular data offloading compared to the baseline traditional infrastructure-only system (a). Offloading through a wireless Access Point (b). Offloading through terminal-to-terminal transmissions (c).

techniques (e.g., LTE-D2D).[1] Besides the obvious benefit of relieving the infrastructure network load, shifting data to a complementary wireless technology leads to a number of other improvements, including: the increase of the overall throughput, the reduction of content delivery time, the extension of network coverage, the increase of network availability, and better energy efficiency. These improvements hit both cellular operators and users; therefore, offloading is often described in the literature as a *win-win* strategy [5]. Unfortunately, this does not come for free, and a number of challenges need to be addressed, mainly related to infrastructure coordination, mobility of users, service continuity, pricing, business models, and lack of standards.

For the reader's convenience, we depict in Fig. 1 the two main approaches to offload in cellular networks when compared with the traditional infrastructure-only mode (Fig. 1(a)). Diverting traffic through fixed WiFi Access Points (AP), as in Fig. 1(b), represents a conventional solution to reduce traffic on cellular networks. End-users located inside a hotspot coverage area (typically much smaller than the one of a cellular macrocell) might use it as a worthwhile alternative to the cellular network when they need to exchange data. Hotspots generally provide better connection speed and throughput than cellular networks [6]. However, coverage is limited and mobility is in general constrained within the cell. Since the monetary cost of deploying an array of fixed APs is far lower than deploying a single cellular base station, the major worldwide cellular providers such as AT&T, Verizon, T-Mobile, Vodafone, and Orange have started integrating an increasing number of wireless APs in their cellular networks to encourage data offloading [7]. Meanwhile, a growing number of applications that automatize the offloading process are proposed for popular mobile devices (mainly iPhone and Android based), such as iPass [8] or BabelTen [9].[2]

The increasing popularity of smart mobile devices proposing several alternative communication options makes it possible to deploy a terminal-to-terminal (T2T) network that relies on direct communication between mobile users, without any need for an infrastructure backbone (Fig. 1(c)). This innovative approach has intrinsic properties that can be employed to offload traffic. T2T-offloading represents a vibrant research topic that we discuss in detail along the survey. Benefiting from shared interests among co-located users, a cellular provider may decide to send popular content only to a small subset of users via the cellular network, and let these users spread the information through T2T communications and opportunistic contacts. Note also that these two forms of offloading (AP and T2T based) may be employed concurrently, enabling users to retrieve data in a hybrid mode.

Although mobile data offloading can be – at a very high level – categorized according to these two classes (i.e., using fixed hot spots or T2T transmissions between mobile nodes), a more refined classification of offloading techniques is required to provide a comprehensive picture. Thus, the main contributions of this survey are three-fold:

- To categorize existing techniques based on their requirements in terms of content delivery guarantee and summarize previously published works.
- To describe a general architecture to enable mobile data offloading with tight or loose delay guarantees.
- To discuss open research and implementation issues.

To the best of our knowledge, it exists up to now only one work from Aijaz et al. that summarizes existing mobile data offloading techniques, although from a higher level and business-oriented point of view [11].

The rest of the survey is structured as follows. In Section II, we propose a general classification of the available mobile data

[1]However, for completeness, we will briefly review other strategies such as femtocells, cognitive offloading, and multicast in Section VII.

[2]The ability to switch seamlessly between heterogeneous networks is referred to as vertical handover [10].

TABLE I: A classification of mobile data offloading strategies, along with their research directions and surveyed works.

Strategy	Delay Requirements: Non-delayed	Delay Requirements: Delayed
AP-based	AP Deployment and Modeling [12], [13], [14], [15], [16], [17], [18], [19]. 3GPP Standardization [28], [29], [30], [31], [32], [33]. Transport Protocols [40], [41], [42], [43].	Prediction-Based Offloading [20], [21], [5], [22], [23], [24], [25], [26], [27]. Feasibility and AP Deployment [12], [34], [5], [35], [36], [37], [38], [39].
T2T	Cooperative Distribution [44], [45], [46], [47], [54], [55], [56], [57], [58], [59], [60], [61]. D2D Capabilities Integration [70], [71], [72], [73], [74], [75], [76], [77], [78], [79], [80].	Subset Selection [48], [49], [50], [51], [52], [53]. Architecture [62], [63], [64], [65], [66], [67], [68], [69].

offloading techniques. In Sections III and IV, we discuss the state of the art for each category, offering also an in-depth literature review. In Section V, we present a reference architecture and its possible implementation into a real framework. In Section VI, we discuss the performance evaluation aspects of different offloading strategies. In Section VII, we examine other possible solutions to the mobile data explosion problem. Finally, in Section VIII we discuss open research challenges, concluding the paper in Section IX.

II. Classification

We may find in the literature various offloading strategies. In this section, we review the main strategies and provide a comprehensive categorization of existing solutions. It is important to pinpoint that mobile data offloading techniques can be classified depending on the assumptions one can make on the level of synergy between cellular and unlicensed wireless networks, as well as the involvement of user terminals in the offloading process. Beyond the distinction between AP-based and T2T approaches, another aspect plays a major role in the categorization. In particular, we take into consideration the requirements of the applications generating the traffic in terms of delivery guarantees. For this reason, we also consider a temporal dimension in the classification, depending on the delay that the data we want to offload may tolerate upon delivery. This translates into two additional categories: (i) non-delayed offloading and (ii) delayed offloading.

We consider these two orthogonal dimensions (delivery delay guarantees and offloading approach), which correspond to four possible combinations, as shown in Table I. The biggest difference between non-delayed and delayed offloading mechanisms lies in the way the timeliness of content reception is handled. In fact, in non-delayed offloading we do not have any *extra* delay on the "secondary" interface (considering cellular the "primary"), while in delayed offloading the network adds some delay (either associated to the fact that the user has to wait until it gets close enough to a WiFi AP, or to get messages through opportunistic contacts).

Non-delayed offloading. In non-delayed offloading, each packet presents a hard delivery delay constraint defined by the application, which in general is independent of the network. No extra delay is added to data reception in order to preserve QoS requirements (other than the delay due to packet processing, physical transmission, and radio access). For instance, interactive audio and video streams cannot sustain any additional delay in order to preserve their real-time requirements. One has to consider that tolerable latency for voice connections is around 50 ms (up to one second for live video streaming). This requirement puts a strain on the network that should meet this deadline to ensure the proper functioning of the application. It turns out that non-delayed offloading is essentially unfeasible in opportunistic networks, since the accumulated end-to-end delay over the transmission path may be too high with respect to the strict delivery requirements. However, if we restrict the analysis to low mobility scenarios, it is still possible to deliver data with strict delay guarantees using T2T transmissions or with the aid of a fixed infrastructure. Non-delayed offloading in most cases may be difficult to implement if one considers that users are mobile and able to switch between various access technologies. If operators want to allow users to be truly mobile and not only nomadic inside the coverage area, they should focus on issues such as transparent handover and interoperability between the alternative access technologies and the existing cellular infrastructure. For instance, this aspect is not granted when one considers a basic offloading implementation through IEEE 802.11 APs. On the other hand, this commitment allows offloading data such as voice over IP (VoIP) or interactive applications, obtaining a nearly transparent offloading process.

Delayed offloading. In delayed offloading, content reception may be intentionally deferred up to a certain point in time,

in order to reach more favorable delivery conditions. We include in this category the following types of traffic: (i) traffic with loose QoS guarantees on a per-content basis (meaning that individual packets can be delayed, but the entire content must reach the user within a given deadline) and (ii) truly delay-tolerant traffic (possibly without any delay guarantees). The relaxation in the delivery constraint allows also moving traffic opportunistically, which, by definition, can only guarantee a probabilistic delivery time. If data transfer does not end by the expected deadline, the cellular channel is employed as a fall-back means to complete the transfer, guaranteeing a minimal QoS. Despite the loss of the real-time support due to the added transmission delay, note that many mobile applications generate content intrinsically delay-tolerant – just think about smartphone-based applications that synchronize emails or podcasts in background. Enabling an alternate distribution method for this content during peak-times (when the cellular network is overloaded or even in outage) becomes an interesting extension and represents a fundamental challenge for offloading solutions.

Eventually, the categorization proposed in Table I may also take into account additional parameters, such as the role of mobility in the process. Delayed offloading strategies rely so much on mobility that we can regard it as a real enabler. Thanks to mobility, users may reach an IEEE 802.11 AP or a neighbor that carries the content of interest, increasing the offload capacity. On the other hand, in non-delayed offloading, mobility often represents a major obstacle and requires a substantial effort in order to make things work together.

III. Non-delayed Offloading

Non-delayed offloading is the most straightforward and experimented class of offloading. Data may be real-time and interactive, thereby enabling the fruition of services such as video streaming and VoIP. So far, WiFi hot-spots have represented the most logical solution due to their widespread diffusion, acceptable performance, and low cost. Operators can incentivize subscribers to offload by offering unlimited data through WiFi hot-spots (and leveraging instead on the capped cellular data). Nevertheless, we can find in the literature many approaches that exploit T2T content sharing between neighboring nodes.

From a technical point of view, cellular base stations are designed to cover large macro areas (1-2 km of diameter in urban areas, and 4-12 km in rural areas), while IEEE 802.11 standard covers limited areas, in the 30-100 meters range. In contrast, the transmission rate is usually much faster for wireless local area networks than cellular technologies. For instance, LTE can reach a shared 28 Mbit/s peak in favorable conditions, with a more realistic average of 10 Mbit/s [6]. On the other hand, IEEE 802.11 standard with its latest amendment can reach a realistic shared throughput of 40 Mbit/s [6].[3] Therefore, as suggested in some works, it is possible to take advantage of this complementarity, combining properly the strengths of different technologies [12], [13].

[3]The advertised throughput is around 100 Mbit/s for LTE and 300 Mbit/s for IEEE 802.11n.

Proposals to employ T2T offloading make use instead of a multitude of wireless technologies. An additional classification could divide T2T approaches in two extra categories: (i) solutions that rely on alternative unlicensed communication technologies to establish direct communications (out-of-band) and (ii) solutions that dedicate part of the licensed cellular band to T2T communications (in-band). In the out-of-band category, IEEE 802.11 and Bluetooth are common choices since they are the most popular wireless technologies present on smart mobile devices today. Other works propose alternative wireless technologies, such as IEEE 802.15, or other less known high-speed short-range communication medium, such as TransferJet [81], WiGig [82] or FlashLinQ [83]. For the in-band offloading category, recent developments of the 3GPP LTE-Advance standard (Rel-12) propose to integrate T2T communication capabilities into future cellular architectures (better known as device-to-device D2D) [84]. However, to date, T2T technologies using unlicensed band (like WiFi and Bluetooth) are the only realistic candidates for data offloading. This because the standardization of in-band D2D communications as an *underlay* to a cellular network is still in its early stages, with a time to market expected in several years [60].

A. AP-based

The prevailing AP-based offloading model today is *user-driven*, meaning that users must explicitly enable the alternative access network in order to benefit from an enhanced experience.[4] This approach is appealing at first, as it requires no modifications in the network infrastructure; however, common limitations such as constrained mobility and lack of session continuity hinder its mass adoption. To pave the way for better cross-resource utilization and improved customer experience, the current trend is to let operators have a deeper control of the offloading process. This eventually raises the question of how a cellular operator can run a profitable business by shifting off-network large parts of its traffic.

Providers are more and more looking toward a tighter integration of alternative access networks and their cellular infrastructure, as depicted in Fig. 2. The integration process concerns partnerships between cellular and wireless providers, common billing and accounting policies, shared subscriber databases for authentication, authorization, accounting (AAA), and security provisioning. Two possible network architectures to date are envisioned to integrate cellular and WiFi access: *loose coupling* and *tight coupling*. In loose coupling, the two networks are independent and are interconnected indirectly through an external IP network. Service continuity is provided by roaming between the two networks. In tight coupling instead, the two networks share a common core and many functions, such as vertical and horizontal handover, integrated management of resources, and common AAA.

1) ***AP Deployment and Modeling****:* Several trace-based analyses demonstrate that the deployment of fixed APs is a viable method to reduce congestion in cellular networks. These

[4]Smart mobile devices already give priority by default to WiFi when a wireless network results available and WiFi interface is enabled.

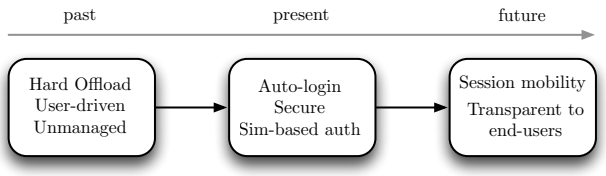


Fig. 2. Steps toward the integration of alternative access networks and the cellular infrastructure.

studies motivate the increasing interest in data offloading, providing an experimental upper bound on how much data it is possible to offload given an existing AP deployment. An interesting strategy to boost offloading performance is to place optimally the APs in order to maximize the traffic that flows through the alternative channel. Since the optimal positioning problem becomes quickly intractable (NP-hard) as the amount of APs increase, a number of sub-optimal algorithms have been developed to this extent.

Lee et al. present one of the first quantitative studies on the benefit that offloading data through APs may bring to network providers and end-users [12]. The availability of WiFi is analyzed during two and a half weeks in Seoul, Korea. The collected traces reveal that, today, non-delayed offloading could relieve a large portion of cellular traffic. These results are further strengthened by Fuxjager et al., which provide records for the Viennese urban district [13]. These studies disclose that, in metropolitan areas, the availability of WiFi APs is already high. Although these results confirm the potential of data offloading as a viable solution for the cellular overloading problem, we should be aware of the limitations of this type of analysis. One of the primary reasons is that the wireless interface in smart phones is less performant than in laptops, due to a number of physical constraints, resulting in lower offloading performance [14]. In addition, measurements based only on signal strength do not consider the issues related to higher-layer protocols, which may influence the theoretical possibility of offloading as perceived from a pure signal strength analysis.

Hu et al. focus on the QoS improvement that non-delayed offloading brings as a function of the number of APs [15]. Simulations, performed with an accurate radio propagation model in an urban scenario, disclose a linear increase in the average throughput per user, as the density of APs increases. The focus is on indoor traffic; thus, the mobility of users is not taken into account in the evaluation. By deploying 10 APs/km^2, the average user throughput can increase by 300% while the number of users in outage decreases by 15% compared to the base case where only cellular networks are present. Three AP deployment algorithms are proposed: traffic-centric, outage-centric, and random uniform. The traffic-centric algorithm aims at increasing the average throughput of the network, while the outage-centric algorithm improves the network outage. In the same spirit, Ristanovic et al. [16] and Bulut and Szymanski [17], come up with AP deployment algorithms aiming at maximizing the fraction of offloaded traffic. All these approaches are similar, proposing a heuristic solution to the problem of the optimal AP deployment, which is inherently NP-hard. The idea is to place the APs close to the locations with the highest density of mobile data requests (or the number of users). Simulation results show that it is possible to shrink cellular traffic by $20-70\%$, depending on the AP density.

Besides simulation results, analytical models help to derive theoretical bounds on performance. Mehmeti and Spyropoulos propose a model based on queuing theory to understand the performance of AP-based offloading with real-time data from the user perspective [18]. They obtain a closed form expression for the expected delivery delay as a function of the AP availability, the traffic intensity, and the rate of the two networks. Similarly, Singh et al. model offloading considering the eventuality of congestion at the APs [19]. The resulting strategy maximizes the number of users capable of reaching their maximum data rates by taking into account the received SINR and the spatial distribution of users and APs.

Discussion. Optimal AP deployment might be a short-term solution for improving performance of real-time data offloading. Ideally, up to 70% of traffic could be offloaded through a carefully planned deployment. Indeed, if the pattern of requests changes, the selected deployment might not be optimal anymore. Furthermore, all reviewed works assume perfect vertical handover mechanisms, which is an over-simplification. Counter-intuitively, adding too many APs could worsen the situation due to mutual interference. An interesting future research area concerns the selection of the optimal AP when multiple APs are simultaneously available. This shares some similarities with the problem of deciding which terminal and which traffic flow to move to a different communication channel [85]. Furthermore, it is related to the Access Network Discovery and Selection Function (ANDSF) mechanism introduced later. On the other hand, analytical models help understand the optimal fraction of data to shift on the alternate channel to maximize the overall data rate and the amount of cellular savings.

*2) **3GPP Standardization Efforts**:* The LTE network proposes an Evolved Packet Core (EPC) flat architecture, fulfilling the requirements for an integrated hybrid network. The EPC is an access-independent all-IP based architecture, capable of providing the handover between IP-based services across a broad range of access technologies (e.g., cellular, WiFi, and WiMAX). Both 3rd-Generation Partnership Project (3GPP) radio access networks and non-3GPP technologies are supported. 3GPP considers data offloading as a key option to tackle the cellular overload problem, proposing the ANDSF mechanism to trigger the handoff between different access technologies [86]. It also proposes three alternative offloading mechanisms that take advantage of the hybrid architecture of the EPC: Local IP access (LIPA), selected IP traffic offload (SIPTO) [87], and IP Flow Mobility (IFOM) [88].

ANDSF is a framework for communicating to the mobile devices the policies for network selection and traffic routing, assisting them in the discovery and handover process [28]. Three different access selection strategies are evaluated, based on coverage, SNR, and system load. A congestion control mechanism to assist ANDSF is proposed by Kwon et al. [29]. LIPA is part of the femtocell architecture and allows a mobile

Binding Cache

HoA	Binding-ID	CoA	Priority
PrefH::MN	BID1	Pref1::MN	5
PrefH::MN	BID2	Pref2::MN	15

Flow Bindings

Priority	Flow-ID	Traffic Selector	Binding-ID
10	FID1	srcAddr = CN1	BID1
20	FID2	srcAddr = CN2	BID2
30	FID3	TCP	BID3

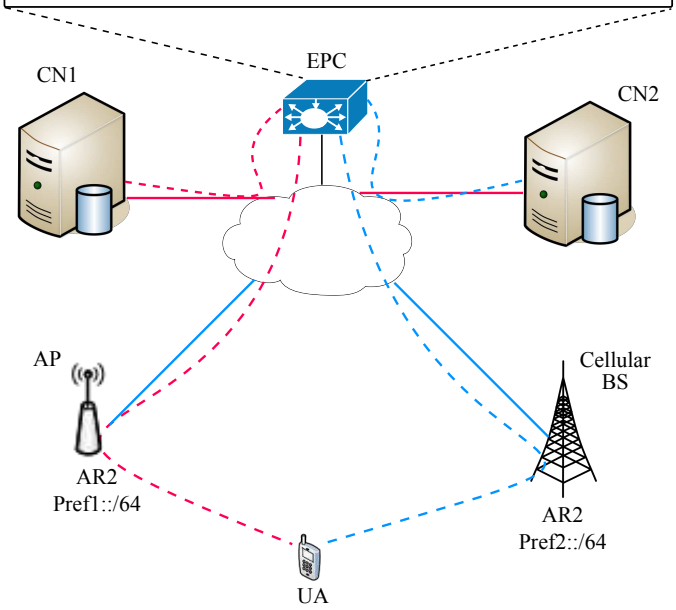


Fig. 3. IP flow mobility (IFOM) procedures: UA has two different CoA, but is bound to the same HoA, allowing the EPC to route different data-flows to alternative access networks. Data coming from CN1 will always transit through WiFi, while data from CN2 is routed on the cellular network. TCP traffic is always forwarded via the interface with the highest priority (WiFi in this case). The bindings between IP flows and entries in the binding cache are stored in the EPC.

terminal to transfer data directly to a local device connected to the same cell without passing through the cellular access network. SIPTO, instead, attempts to offload the core of the network, balancing data flows to selected IP gateways at core level. Note that these solutions (e.g., LIPA/SIPTO) offload the core cellular network and do not relieve bandwidth crunch in the access network. Therefore, they are not the focus of this survey. We suggest interested readers to refer to Samdanis et al. [89] and Sankaran [90].

IP Flow Mobility (IFOM) implements offloading at RAN level, allowing providers to move selected IP data-flows between different access technologies without disrupting ongoing communications [30]. Conversely to ANDSF, which is utilized to discover, connect, and manage handover between neighboring APs, IFOM provides offloading capabilities in terms of moving data-flows between access networks. IFOM allows terminals to bind multiple local addresses (CoAs) to a single permanent home IP address (HoA), and to bind distinct IP flows (e.g., HTTP, Video, VoIP) to different CoA, as depicted in Fig. 3. This feature allows different flows related to the same connection to be routed over different radio access technologies based on some operator-defined policy. Sometimes IFOM involves a total switchover of all traffic from one access technology to another. In other cases, the network allocates only "best effort" data to the complementary access, while keeping delay-sensitive flows on the cellular network. IFOM allows users benefiting from high bandwidth connections when at least one complementary network is available. At the same time, operators are able to manage the radio access resources optimally, reducing the network overload and providing different QoS levels to distinct data-flows. Drawbacks of IFOM reside in the additional modifications needed both at terminal and network levels to manage the heterogeneity of access technologies. In addition, in very dense wireless environments the management of user mobility should adapt to very challenging conditions, such as interference and dynamic terminal reconfiguration. A significant problem is that QoS-based routing will hardly work in the case of encrypted flows such as IPsec or SSL/TLS. Only when WiFi will be considered a trusted access technology by the 3GPP, operators will be able to apply QoS traffic reclassification at access network level [31]. In an attempt to improve performance further, Makaya et al. suggest to integrate IFOM with a multilink striping manager capable of distributing the same data flow across different radio interfaces, according to application and network status [32]. The striping manager employs periodical reports on link quality and network congestion to determine on which interface to send data. The striping manager inspects all the packets in order to assign them to a specific flow. Testbed results show that the aggregated throughput improves by 20%, ensuring also a seamless support to service continuity in case of link degradation.

Nevertheless, it is perfectly legitimate to wonder whether it is a good idea to devise 3GPP-specific solutions outside of TCP/IP, and how widely this specific solution would be implemented in future. For these reasons, Korhonen et al. present three different fully IP-compliant offloading solutions, not requiring any specific access technology integration [33]. Backward compatibility is assured without relying on any system specific extension. The first proposed solution allows the network to push new routes and policies to the UE through a DHCPv6 protocol exchange; the second method is developed on top of the IPv6 neighbor discovery protocol (RFC 4191 [91]), and exploits the possibility given by this protocol to remotely control the default router (in this case the default interface) for different data flows; the third approach extends the second solution by adding IPv4 capabilities. Table II proposes a summary of existing solutions. *IPv4/IPv6* indicates to which IP version the offloading strategy applies. *Dynamic* characterizes whether the offloading strategy can be updated during the ongoing session. *Direction* indicates who initiates the offloading procedure (operator or user-initiated). *Offload* defines the offloading targets. The first four strategies are 3GPP-specific, while the last three are fully IP-compliant.[5]

Discussion. 3GPP standardized the ability to perform of-

[5]Note that other offloading solutions have been proposed by 3GPP, such as Multiple access PDN connection (MAPCON), which is a subset of IFOM, and S2A MObility based on GTP (SaMOG), which does not guarantee address preservation and works only with trusted non-3GPP networks.

TABLE II: 3GPP vs. non-3GPP offloading solutions.

Strategy	IPv4/v6	Dynamic	Direction	Offload
ANDSF Only	Both	No	Operator	Access/Core
SIPTO	Both	No	N/A	Core
LIPA	Both	No	Operator	Access/Core
IFOM	Both	Yes	Operator/UA	Access
DHCPv6	IPv6	UE inits	UA	Access/Core
RFC 4191	IPv6	Yes	Operator	Access/Core
RFC 4191 + IPv4	Both	Yes	Operator	Access

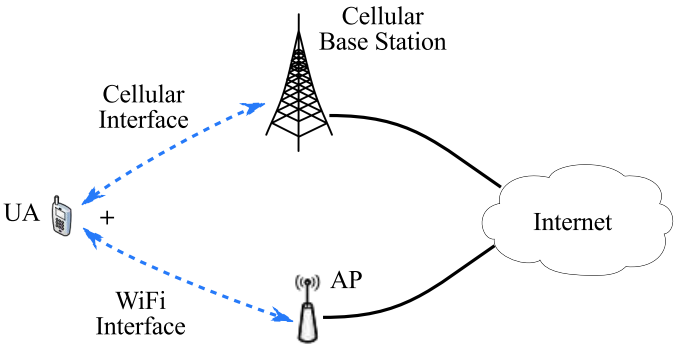


Fig. 4. Multiple interfaces can be exploited simultaneously by users to increase the throughput, and improve data coverage. Transmission protocols should be capable of handling efficiently such situations.

floading through a variety of access methods in the LTE network architecture. New protocols, such ANDSF and IFOM, transform offloading into a nearly transparent mechanism for end-users. Operators are able to shift selected data-flows between different access technologies without any disruption. This concurs in lowering the network congestion. As of today, no commercial deployments of ANDSF and IFOM exist, though trials are undergoing to understand the feasibility of these solutions. The widespread adoption of these techniques is one of the keys to enable effective operator-driven offloading strategies. The mechanisms presented in this section are, as today, the standard frameworks in which forthcoming AP-based offloading strategies need to be integrated. On the other hand, these solutions could be significantly improved by considering delayed reception and opportunistic transmissions.

*3) **Multi-Interface Integration and Transport Protocols**:* The development of a novel IP-based transport protocol is an essential prerequisite to enable future offloading capabilities to mobile smart devices. This new transport protocol should be able to cope with seamless switch overs, different simultaneous connections and aggregation between multiple access technologies, as explained in Fig. 4. These functionalities cannot be implemented on top of current standard Internet protocols, so we must consider extensions to existing ones.

Considering a vehicular scenario, where offloading is challenging due to the mobility of users and the limited transmission range of WiFi APs, Hou et al. design a novel transport protocol that exploits the potential of the opportunistic use of complementary access technologies, striping and transmitting data across multiple network interfaces at the same time [40].

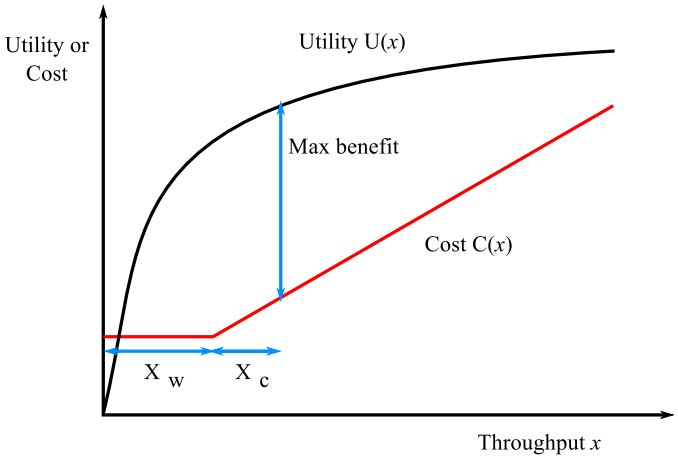


Fig. 5. Plot of utility versus cost and benefit functions from [40]. X_w and X_c are the instantaneous throughput on the WiFi and the cellular interface. The cost C(x) of the WiFi interface is assumed constant, while the cost on the cellular interface depends linearly on the throughput.

In order to calculate the right amount of data to be transmitted on each interface, the proposed system models the user utility, trading off between throughput and connection cost, within an optimization framework that maximizes the cost-utility benefit, as shown in Fig. 5. The scheduling logic is implemented above SCTP (Stream Control Transmission Protocol) [92], which natively can bind multiple IP addresses at each communication endpoint. The proposed framework adds striping and throttling capabilities to the standard SCTP implementation. Real world experiments claim a $65-80\%$ cellular data reduction.

Patino et al. consider instead the implementation of a multipath transmission protocol (MPTCP) to use simultaneously several networks to transmit [41]. MPTCP is emerging as a valuable option to provide multiple connectivity over different interfaces [93]. An additional advantage of MPTCP is that it does not need any additional requirements on the network side, being entirely implemented at end-hosts. Nevertheless, it is possible to employ a MPTCP proxy, adding the possibility to communicate with a correspondent non-MPTCP enabled device, as depicted in Fig. 6. MPTCP has several working implementations, notably on Android smartphones [94], and a large scale commercial deployment inside Apple iOS 7 operating system [95].

Limiting themselves to the problem of switching seamlessly between several available access technologies, Nirjon et al. present MultiNets, a seamless client-based transport layer capable of dynamic handover between different network interfaces [42]. Since it is impractical to jump from one network interface to another with connection-oriented data-flows, MultiNets allocates new connections on the new interface, waiting for pre-existing ongoing TCP-like sessions to terminate naturally, before shutting down the old network interface in order to prevent any data-flow disruption. MultiNets is implemented as a transport layer solution that any application can access through the exposed APIs. MultiNets enforces three different strategies, which impact the handover preemptiveness: energy savings, offload, and maximum throughput. Extending the same concept, Rahmati et al. discuss how to move the ongoing TCP flows between different network interfaces seamlessly,

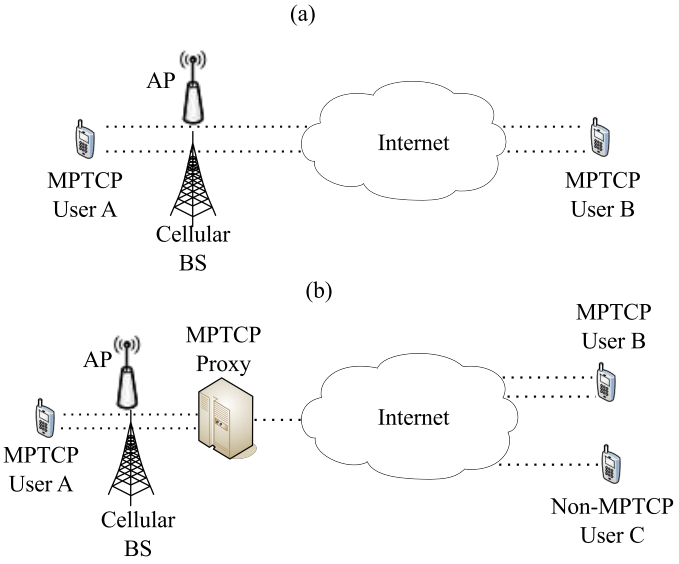


Fig. 6. MPTCP deployment from [41]: (a) MPTCP is used for end-to-end transmission: MPTCP-capable clients are needed at both ends; (b) an MPTCP proxy is introduced in the network: MPTCP-capable devices can communicate also with non-MPTCP hosts.

without any modification in application, infrastructure, or existing protocol behaviors [43]. They implement the Wait-n-Migrate mechanism to optimize interface switching in case of short lived data-flows. Each time a new access network becomes available, Wait-n-Migrate assigns to ongoing data-flows a wait-time before being switched. In addition, they propose a "resumption agent" method to integrate the Wait-n-Migrate strategy, leveraging on the resume function support of some applications. These two methods, if used concurrently, could mitigate the impact of the varying network conditions.

Discussion. The transition toward the simultaneous use of multiple access technologies brings a number of issues. In order to benefit the most from non-delayed offloading, it becomes mandatory to develop innovative communication stacks beyond classic IP-protocol, capable of supporting advanced features (e.g., multiple instantaneous connections, data aggregation, and inter-technology switchovers). Extensions to standard protocols started to appear to cope with these issues (e.g., SCTP, MPTCP), enabling to aggregate together the bandwidth offered by different technologies, and allowing seamless handover between distinct access technologies. Nevertheless, a widely accepted transport protocol to handle transparently several flows in parallel on separate interfaces has not yet been standardized.

B. T2T

Real-time T2T offloading is often associated with cooperative strategies to exploit concurrently the availability of multiple interfaces. Thus, the network should be capable of coordinating data retrieval in a distributed fashion. Initially, only out-of-band transmissions were considered. However, the latest developments in the 3GPP LTE Standard (Rel-12) propose integrating direct in-band communication capabilities into the future cellular architecture [84]. This provides additional flexibility to the network but raises issues such as mutual interference and resource allocation, since T2T transmissions take place in the same band as the cellular transmissions. Cooperative data retrieval is shown to improve the spectral efficiency of the network [96]. While classical studies show that the theoretical transport capacity of multi-hop ad hoc networks scales sub-linearly as $\Theta(\sqrt{n})$ [97], [98] with the number n of users, cooperation among nodes brings linear scaling law of $\Theta(n)$ [99]. Besides improvements in terms of congestion, real-time offloading techniques through T2T communications offer advantages if compared to standard cellular distribution, in terms of average throughput, coverage, and energy consumption, at the cost of higher complexity.

If peers are not stationary, link quality may suddenly change, making it difficult to guarantee QoS. If data delivery can be deferred, a better candidate for data distribution is delayed offloading (see Section IV-B). To guarantee real-time requirements, most architectures assume low mobility and co-located peers interested in receiving a common content [61].

*1) **Cooperative Data Distribution**:* Kang et al. propose CHUM, a turn-based download strategy, where the designated proxy downloads multimedia content through the cellular network and then multicasts it on the WiFi interface to other interested nodes [44]. This method aims at cutting the connection costs up to 90%, while maintaining fair resource usage. An extension has been developed to cut battery depletion in the case of IM (Instant Messaging) service [45]. Simulation and testbed results show that energy savings increase with the number of cooperating nodes. Similarly, the COSMOS architecture exploits the availability of an alternative channel to broadcast a real-time stream to nearby nodes [46]. Like a peer-to-peer network, COSMOS is resilient to node failures. The number of successive broadcasters is tuned following the density of nodes involved in data distribution, to trade off collisions on the wireless medium and data redundancy.

Stiemerling and Kiesel consider cooperative T2T streaming in high mobility scenarios [47]. The basic idea is to download each video chunk only once through the cellular channel, and to share it through short-range links. Cellular accesses have to be coordinated among willing nodes in order to relieve the cellular infrastructure. One node acts as the central controller, and estimates the available throughput on the cellular link for each other node, so to coordinate content retrieval among peers. Simulation results provide the minimum number of cooperating nodes required to achieve the target values of throughput and reception delay. Karunakaran et al. consider the variable data rate that cellular users can reach, to transmit data only to those users with the best channel quality [54]. Data is subsequently relayed to all other nodes by means of T2T transmissions. Compared to cellular-only distribution, a rate-proportional scheme performs better in terms of energy consumption, reduced by 70%, and average throughput, increased by a factor 2.

Hua et al. present a scalable video multicast solution that jointly exploits cellular broadcast, video coding and T2T transmissions [55]. The base video layer is broadcasted to all the users within the cell. The enhancement layers, instead, are transmitted only to a subset of users. Modulation and

coding schemes employed for transmission are the outcome of a joint optimization problem involving T2T transmissions and cellular coverage. The enhancement layers are then forwarded to remaining users through T2T transmissions. Simulation results show heavy increases in terms of PSNR. Seferoglu et al. include network coding and broadcast in the optimization problem to reach the transmission rate at the source that maximizes the average user throughput [56]. A working version of this approach, named MicroCast, is presented by Keller et al. [57]. Another testbed for cooperative real-time video streaming among mobile nodes is proposed in [58]. Experimental results on power consumption show that the proposed approach is beneficial from the user point of view. A theoretical model for power consumption is developed to support experimental results. A cluster based content distribution system is implemented in [59].

Finally, Andreev et al. overcome the classical problems of cooperative offloading (i.e., neighbor discovery, connection establishment, and service continuity) by adopting a network driven approach [60]. In their proposal, the cellular network architecture is intelligent enough to assist connected users in the content discovery and connection establishment phases. Simulation results show a 2.5 times boost in throughput, offloading around 30% of the traffic.

Discussion. The complexity of cooperative content distribution is high, involving the joint optimization of different access technologies, interference, transmission rates, scheduling and energy efficiency. Centralized or distributed solutions have been developed and tested through simulation, theoretical analysis and real test beds. Optimal solutions are NP-hard, so heuristics need to be adopted. Most of the papers focus on how to achieve enhanced data rates, saving at the same time battery. In this context, an energy consumption model is provided in [61]. Security and trust considerations concur in making the problem even more complex. A novel approach, involving a continuous control wielded by the network, could possibly simplify the problem. It is the focus of the following section.

*2) **Device-to-Device Capabilities Integration**:* Recent developments in the 3GPP LTE Standard (Rel-12) propose integrating direct in-band communication capabilities into future cellular architectures [84], often also referred to as cellular network underlay [70] or device-to-device (D2D), rather than using traditional technologies working on unlicensed bands (mainly IEEE 802.11 and Bluetooth). This paves the way for a combined use of cellular and short-range transmissions, offering users various degrees of freedom for transmission and a network-assisted environment. End-users discover each other in proximity through explicit probing [70] or via the access network guidance [71]. Additional discovery options are examined in [72]. Upon discovery, nodes can communicate using either dedicated resources or a shared uplink cellular channel [73]. D2D communications are then triggered by the cellular network, and fall under continuous network management and control. For these reasons, they can also be employed for load balancing purposes [100]. Hence, D2D could become the ideal platform to develop data offloading in the future, because it may achieve higher resource utilization by reusing the spectrum of physically neighboring devices, while reliability may significantly increase thanks to a shorter link distance. Furthermore, D2D capabilities enable LTE to become a very interesting technology for public safety networks [101]. Anyway, critical issues such as neighbor discovery, transmission scheduling, resource allocation and interference management, in particular in the case of multiple cell deployments, still need to be addressed in order to proceed to the effective integration in future cellular architectures. Related tutorials provide the reader with a broader overview on the existing research challenges and applications of D2D [102], [103].

Interference management and transmission coordination represent thorny problems that must not jeopardize the QoS of cellular users in the primary network. When two or more pairs of neighboring nodes are willing to communicate, they may use the same resources. In this case, interference is a major issue. The network could limit the maximum transmission power of D2D peers [70]. The optimization of radio resource allocation help decrease the mutual interference between D2D communications and the primary cellular network [74]. Similarly, joint resource allocation and power control schemes can also be adopted [75]. Note that, for the intrinsic real time requirements of cellular networks, the computational complexity of resource allocation algorithm represents a tangible issue [76]. The resource allocation problem is examined in the case of static relay nodes by Hasan et al. [77]. Exploiting a different approach, Li et al. adopt social-networking methods to address the allocation problems [78]. For instance, resources are allocated proportionally to centrality and community rankings. Neighbor discovery intervals depend as well on the centrality of a node.

Li et al. study the realistic bound of an offloading strategy exploiting LTE-D2D in a large-scale scenario [79]. Nodes are divided into downloaders and helpers, which can aid the cellular network to deliver the content to downloaders. The optimal distribution strategy is formulated as an optimization problem and assessed through simulations. By knowing the mobility pattern of users, an upper bound on performance is devised. Simulation results confirm that augmenting the number of users in the cell largely benefits to offloading, increasing its efficiency. In that case, D2D transmissions account for up to 50% of the traffic. On the other hand, Yaacoub et al. formulate content distribution as an optimization problem that takes into account fairness in energy consumption and cellular channel quality [80]. Both unicast and multicast distribution are considered in the analysis. Game theory concepts are employed to solve the problem analytically.

Discussion. T2T communications as an underlay of cellular networks represent a significant leap forward towards the deployment of heterogeneous networks. D2D communications in this case share resources with cellular transmissions, therefore generating mutual interference. Consequently, resource allocation optimization, power control, and device discovery are key topics for the research community. However, the underlay approach does not exploit surplus bandwidth available through complementary technologies, but rather aims at

taking advantage of parts of the LTE spectrum that may be under-utilized. Still, this could be the ideal technology to support the predicted data growth. Cellular operators can make profits on network-assisted D2D communications, supervising at the same time the resource consumption and the QoS of the network, which is difficult in out-of-band offloading techniques.

IV. Delayed Offloading

As mentioned in Section II, delayed offloading adds a non negligible delay to content reception. In general, while it is crucial for end-users to receive content within the deadline, it is not fundamental to receive the entire stream at a fixed rate. Some content may have an explicit delivery delay bound (even though at the level of the whole content), others may be truly delay tolerant. E-mails, news-related information, or podcasts, to name a few, may sustain a certain degree of delay without breaking user satisfaction. These are therefore excellent candidates to be offloaded with loose delivery bounds.

Most of the time, the offloading strategy relies on the cellular network to bootstrap the distribution process (to infect seed users in T2T-based offloading) or to ensure minimal QoE guarantees (fall-back transmissions when the deadline approaches). Before the deadline, the content is preferably delivered through the alternative technology. Unlike the approaches set forth in Section III, delayed offloading directly exploits the mobility of nodes to create communication opportunities. As a side effect, performance heavily hinges upon the mobility pattern of users. A short digression on mobility characteristics is thus necessary to better catch the fundamental properties and inherent limits of delayed offloading.

Since messages are forwarded only during contacts with users or APs, the statistical analysis of such encounters becomes particularly meaningful. First, the time until a new encounter occurs (the inter-contact time) gives an effective indication of the delivery capacity inside the opportunistic network. In addition, when contacts occur, knowing for how long they last (their contact time) would help us to foresee how many pending messages can be forwarded. The distribution of contact times also affects the total delivery capacity when multiple users compete for the same wireless channel, because contacts can be wasted due to contention and scheduling. These properties have been deeply investigated in trace-based studies [104], [105]. Common understanding is that inter-contact and contact times between mobile users often display a power law distribution with an exponential heavy tail. Analogous results hold also for contacts between users and fixed APs [12]. However, as pointed out by Conan et al. [106] and Passarella et al. [107], these results focus on aggregate inter-contact distributions, and are not representative of the network behavior, which instead depends on the properties of individual pairs. An interesting addition to the standard contact and inter-contact analysis is proposed by Tatar et al., which consider in their model an extended notion of contact relationships [108].

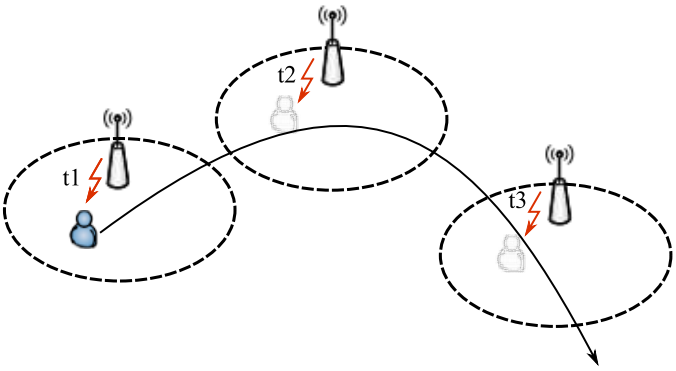


Fig. 7. AP-based offloading. Mobility allows to receive delay-tolerant data from different APs at different times.

A. AP-based

AP-based strategies take advantage of a complementary networking backbone, often formed by fixed WiFi APs, to deliver data bypassing the cellular network. The complementary access network may be part of the cellular operator network, or may be completely separate. In the latter case, an agreement between operators should be envisioned. At first, this approach looks similar to the non-delayed case. The delay-tolerance of content is exploited here, with data exchange happening upon subsequent contacts between the user and different APs exploiting a sort of space-time diversity, as illustrated in Fig. 7. The movement of end-users creates contact opportunities with fixed APs defining the offloading capacity of the network.

Current research efforts aim at predicting the future offloading potential through past behaviors of users such as mobility, contacts with APs, and throughput. Using this prediction, the offloading coordinator may decide which fraction of data to offload, when, and to whom. Possibly, downstream content is split in several pieces, which are then pro-actively sent to APs that nodes will (probably) encounter in the future. An alternative research area aims at identifying the optimal number of fixed APs and their geographical location, starting from a known user's mobility pattern. In the following sections, we present in detail these two approaches.

*1) **Prediction-Based Offloading**:* Siris et al. combine the prediction of node mobility with the knowledge of the geo-localization of fixed APs to enhance the offloading process [20]. The predictor informs the coordinator of how many APs a mobile node will encounter during its route, when they will be encountered, and for how long the user will be in AP's range. The algorithm seeks to maximize the amount of delay-tolerant data to be offloaded to WiFi, ensuring also that data is transferred within its deadline.[6] Similarly, the MobTorrent architecture exploits the hybrid infrastructure, data pre-fetching, and cache replication at fixed APs [21]. Download requests are issued through the cellular channel. Requested data is cached in advance to APs using location information and the mobility history of users.

Dimatteo et al. propose a network-centered architecture called MADNet [5], which integrates cellular, WiFi APs,

[6]In the case of real-time traffic, a slightly modified version of the algorithm is employed in order to reach the maximum available throughput by exploiting the existence of multiple parallel connections.

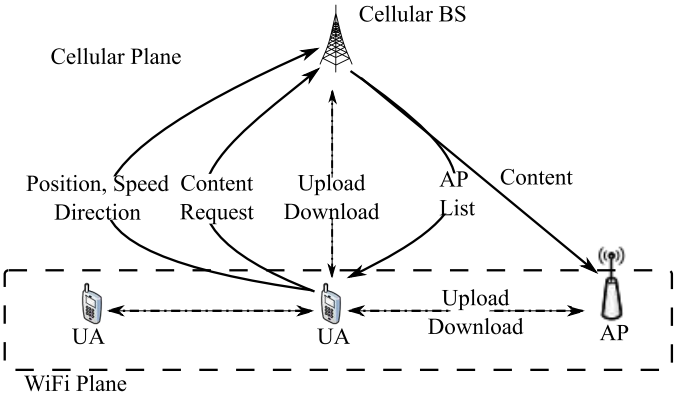


Fig. 8. MADNet system architecture [5]: when a mobile node wants to communicate, it makes a request to the cellular BS, which may replies directly forwarding the content through the cellular network or sending the content to a neighboring AP. The BS predicts the route of the nodes using the status information sent by the mobile node.

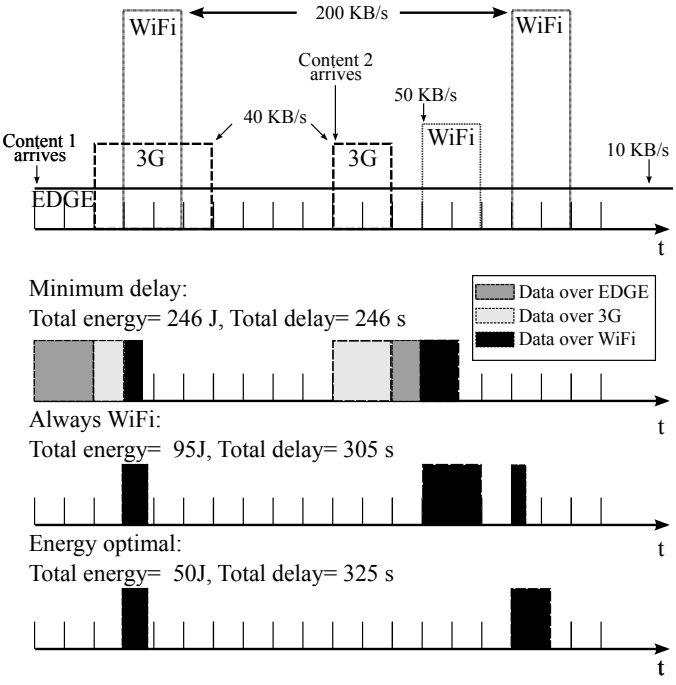


Fig. 9. Three offloading strategies from the SALSA framework [22]: The *Minimum delay* strategy minimizes the total delay, selecting always the channel with the fastest data rate; the *Always WiFi* strategy, uses only WiFi APs, regardless of their data rate; the *Energy optimal* strategy minimizes the energy consumption, using always the most energy efficient channel.

and mobile-to-mobile communications. MADNet employs the cellular network as a control channel. The system is explained in Fig. 8. When a mobile user asks for some content, the offloading coordinator replies with the list of the surrounding APs where it may pick up the requested data. The offloading coordinator predicts the neighboring APs by exploiting positioning information uploaded by users through the control channel. Key results, obtained through simulation, show that a few hundreds APs deployed citywide could offload half of the cellular traffic. Ra et al. discuss the tradeoff between delay and QoS, presenting a centralized optimization algorithm called SALSA [22]. This algorithm determines when deferring a transmission (in case it should be delayed) by adapting the offloading process to network availability and location information. The main contribution of the work is to explore the energy efficiency of delayed transmissions, because WiFi has, in general, better efficiency than cellular transmissions, as depicted in Fig. 9. The transmission decision relies on the prediction of the future available bandwidth for each possible access network, estimated as the average rate achieved over past transmissions, or as a function of the received RSSI. Authors also perform synthetic and real-world experiments to confirm the good performance of SALSA, which can save up to 40% of energy if compared to other baseline offloading strategies. Go et al. suggest a heterogeneous city-scale mobile network that opportunistically offloads some cellular traffic to existing WiFi APs [23]. The core of the system relies on DTP (Delay Tolerant Protocol) to mask network disruptions from the application layer [109]. DTP binds the connection to a unique *flow ID* rather than to a tuple of physical IP addresses and ports, providing to applications the illusion of a continuous connection. The proposed system employs dedicated proxies located at the edge of the access network that hide user disconnections to application servers. Finally, Malandrino et al. relax the assumptions of an accurate prediction scheme by proposing a model that considers the uncertainty of mobility through a Gaussian noise process [24]. Each AP performs a joint pre-fetching and scheduling optimization through a linear programming problem, aimed at maximizing the aggregate data downloaded by users.

Focusing on user-centered policies instead, Balasubramanian et al. design Wiffler [25], an algorithm capable of exploiting the delay tolerance of content and the contacts with fixed APs. Wiffler predicts future encounters with APs, deferring transmission only if this saves cellular traffic, employing the heuristic detailed in Algorithm 1. Wiffler predicts the WiFi transfer size W based on past encounters with APs. The contact history is employed to estimate both the inter-contact time and the average throughput per contact. By means of trace-based simulations, the authors show that with a prediction based only on the last four encounters, Wiffler obtains low prediction error for future intervals of around one minute. The system is able to offload between 20 to 40% of the infrastructure load, depending on the content's delay tolerance. Similarly, Yetim et al. consider the decision of waiting for WiFi encounters rather than using the cellular connectivity as a scheduling problem [26]. Different sizes and deadlines are considered for content. Presuming that each content may be divisible in smaller scheduling units of MTU size, the scheduler exploits short windows of WiFi coverage to shift up to 23% of the total traffic away from the cellular network. Finally, Zhang et al. focus on how to find the optimal instant to hand-back data transfer to cellular networks [27]. Although delayed transfers may substantially improve the offloading performance of cellular networks, delaying all transfers up to their maximum delay tolerance is often an ineffective strategy. In case of absence of WiFi, each delayed transmission frustrates user experience. An ideal solution is to identify the optimal instant of time after which a user should stop deferring transmissions and start transferring data using the cellular interface, trading-off offloading efficiency and user satisfaction. For this reason, the proposed algorithm maximizes

Algorithm 1: Wiffler offloading decision heuristic [25].

$D \leftarrow$ earliest deadline among queued transfers.
$S \leftarrow$ size in bytes to be transferred by D.
$W \leftarrow$ estimated WiFi transfer size.
$c \leftarrow$ tuning parameter.

if *WiFi is available* **then**
  send data on WiFi and update S
end

if $W < S \cdot c$ **and** *3G is available* **then**
  send data on 3G and update S
else
  wait
end

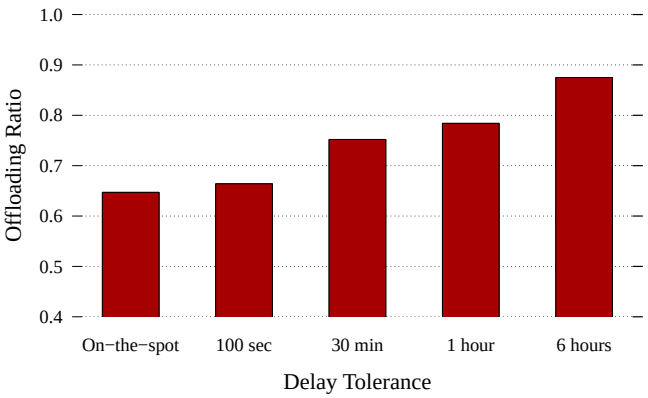


Fig. 10. Offloading ratio of delayed AP transfers with various deadlines, 100% of traffic delayed, Seoul dataset [12]. Increased delay-tolerance values result in an increased fraction of data offloaded.

a utility function depending on both the offloaded data and user satisfaction. The amount of offloaded data is predicted through a contact process modeled as a time-homogeneous semi-Markov process.

Discussion. A key requirement to drive effective AP-based offloading is the ability to predict future capacity. The decision to wait for a possible upcoming offloading opportunity or to transmit data through the cellular channel (considered as a scarce and costly resource) is of utmost importance when dealing with delay-tolerant data. Distributed and centralized prediction methods have been developed based on the knowledge of prior encounters, mobility patterns, AP locations, and bandwidth availability. Future researches in this sense should also take into account the obvious trade-off between the overhead brought by context-awareness and the accuracy of prediction. Furthermore, most existing solutions for AP-based offloading rely on optimization frameworks, which are complex to solve and need heuristics. As a result, an interesting research topic might be to explore alternative self-adaptive approaches (e.g., based on machine learning techniques).

*2) **Feasibility and AP Deployment**:* Similarly to Section III-A1, we address here the feasibility and capacity of AP-based offloading, this time considering delay-tolerant content. Lee et al. demonstrate that increasing the delay-tolerance of content substantially improves the ratio of offloaded traffic, as depicted in Fig. 10 [12]. Additional findings suggest that the average completion time for delayed offloading is always much lower than the maximum deadline. Surprisingly, the authors discover that, with large content, delaying the transmission may result in faster completion times than not delaying it at all. This is motivated by the fact that WiFi usually offers higher data rates than cellular networks, which translate into shorter aggregate completion times. Theoretical bounds for delayed data offloading with WiFi AP are derived analytically by Mehmeti et al. [34]. Mean reception delay and offloading efficiency are evaluated as a function of the number of users and the availability of APs using queuing theory concepts.

Dimatteo et al. [5], Trestian et al. [35], and Lochert et al. [36] discuss optimal placement of APs. The first work quantifies the number of APs required to offer a citywide offloading coverage. The authors argue that, with the integration of only a few hundreds of APs, it may be possible to offload half of the cellular traffic in a metropolitan area. A simple heuristic for optimal AP deployment is proposed. Trestian et al. suggest instead upgrading the network capacity in a limited number of locations, called *Drop Zones* [35]. The underlying intuition is that most users pass by a limited number of hub locations during daily commutes. Thus, by upgrading only a tiny fraction of the network, providers may strategically support growing traffic with minimum investments. The original contribution of this work is the algorithm for *Drop Zones* placement, which aims to reduce both the number of APs and the average uploading delay. This is a minimum set-selection problem, with a NP-hard optimal solution. The paper proposes also a sub-optimal greedy algorithm that guarantees a 24% reduction in APs placement relative to non-delayed strategies. Finally, Lochert et al. propose a genetic algorithm to identify the best AP positions for information dissemination in vehicular networks [36]. Although the work is not oriented toward offloading, the proposed algorithm could substantially contribute in the optimal AP deployment at a large scale.

In the context of vehicular networks, Abdrabou and Zhuang study the minimum number of APs to cover a road segment in order to guarantee a probabilistic connection time [37]. Malandrino et al. model data downloading in a vehicular environment as an optimization problem, considering also the presence of fixed APs [38]. To counter the scarce availability of APs due to placement and maintenance costs, they take also into account parked vehicles, acting as additional APs, to assist in data distribution [39].

Discussion. Similarly to the non-delayed offloading case, performance is tied to AP density. Nevertheless, the time dimension matters here, as increased delay-tolerance translates into an extended fraction of offloaded data. We may find a number of placement algorithms that exploit the delay tolerance of content by adding APs where people are most likely to transit. The problem has similarities with the optimal road-side unit placement strategies for ITS (Intelligent Transportation Systems) applications [110]. Cost based analysis proposed in the literature, help better understand the existing trade off

between the cost of deploying more APs and the offloading benefit [5], [35], [38]; unfortunately, many solutions are not directly comparable due to differences in reference scenarios, use cases, and simulation parameters.

B. T2T

In delayed T2T offloading, content distribution is delegated to end users: in a broad sense, *users are the network*. They actively participate in the dissemination process by exploiting T2T communications. Mobility is an additional transport mechanism, creating opportunities for infected users to transfer data employing a delay-tolerant (DTN) approach [111].[7] DTNs allow content forwarding through store-carry-forward routing regardless of the existence of a connected path between senders and receivers, at the cost of additional reception delays. Golrezaei et al. analyze the theoretical performance bound for throughput [112]. Store-carry-forward routing coupled with simple caching policies at nodes could bring a linear throughput increase in the number of nodes. For these reasons, delayed T2T-based offloading is often seen as a quick and inexpensive way to increase mobile network capacity and to handle the predicted data tsunami [50]. Unlike AP-based approaches, the gain of this schema relies entirely on redundant traffic. However, this proves to be relevant for content access, as popularity follows Zipf-like distributions [113] – a small subset of content results extremely popular and is requested by a large number of co-located users, causing severe congestion and bandwidth shortage at RAN level. Moreover, the DTN approach supports conditions where standard multicast and broadcast approaches (also included in LTE [114]) cannot be used. For example, it supports all cases where popular content is requested by users during a given time window (short enough to guarantee that users are still physically co-located in the same region), but not necessarily at the exact same time. Note however, that DTN-based offloading is also beneficial when multicast in the cellular network can be used [115].

From its characteristics, it follows that the DTN approach can only address the diffusion of data with loose delivery constraints. Content is ideally supplied only to a small fraction of selected users among those who requested it. These *seeds* bootstrap the propagation by transferring content to users within their transmission range, as in Fig. 11. In this category, we also include strategies where the communication opportunities between nodes arise as a side effect of duty cycling of ad hoc interfaces. T2T interfaces are typically energy-hungry, and it is possible to apply energy saving policies to them, dynamically toggling between on and off states [116], [117].

A number of strategies can be used to disseminate the content among mobile nodes. In principle, any forwarding or data dissemination scheme proposed for opportunistic networks can be used. Hereafter, we just give a few examples. Interested readers can refer to [118], [119] for dedicated surveys. From the seminal work of Vahdat and Becker that firstly proposed

[7]Delay-tolerant, disruption-tolerant, opportunistic, challenged, and intermittently-connected networks are used in the literature most of the time as synonyms, although sometimes they denote slightly different concepts. With respect to the offloading solutions, they can be considered as synonyms.

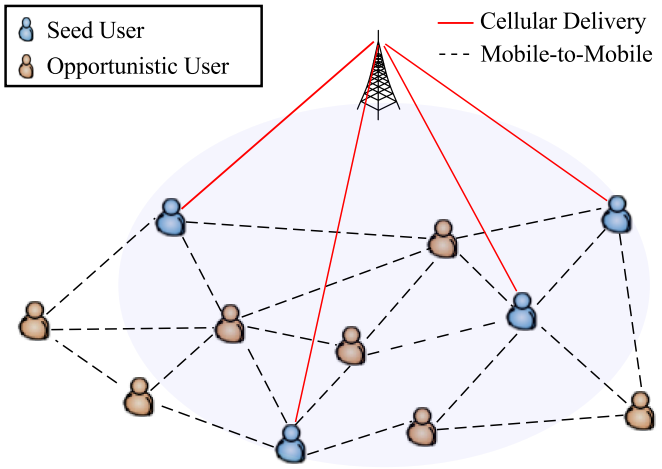


Fig. 11. Data offloading through delay-tolerant networks. *Seed* users initially receive the content through the cellular network. Direct ad hoc transmissions are used to propagate the content in the network

mobility-assisted epidemic forwarding [120], many routing protocols in the context of DTNs have been proposed. Notable works on forwarding strategies from Spyropoulos et al. [121], Lindgren et al. [122], and Burgess et al. [123] go beyond simple epidemics by tackling statistical and mobility characteristics of nodes, and targeting the case of separate subsets of users with different interests. Mathematical frameworks based on ODEs and Markovian models provide theoretical bounds on the performance of dissemination delay and the number of copies of the message in the network [124], [125]. Similarly, analytical bounds on dissemination delays are derived from the speed and density of nodes in [126], [127].

To motivate the utility of DTN-based offloading, Vukadinovic and Karlsson propose a system specially designed for podcast distribution [128]. Podcasts are the ideal content type for DTN-based offloading, for their popularity and delay-tolerance. Consider what happens if, in place of deploying more infrastructure in order to satisfy all the request for content, only best connected users are employed as seeds, receiving podcast directly from the infrastructure. In the envisioned system, the remaining subscribers may collect missing content only upon opportunistic encounters with subscribers of the same feed. Results demonstrate that opportunistic content distribution is a resource-efficient method to increase the spectral efficiency and the aggregate throughput of the network, at a lower cost than deploying additional infrastructure. Optimal seed selection, together with the effectiveness of the DTN diffusion, may entail further significant improvements in RAN overload.

For the reasons listed above, most of the research efforts in this field focus on the design of efficient algorithms for the optimal selection of seed users, in order to minimize the number of users that receive the content through the cellular interface. On the other hand, a number of works deal with network architecture and protocol design. The former approach relies on social networking analysis or machine learning techniques to predict *which* users are the best gateways for content. The latter tackles the choice of *what* type of traffic to offload and *how*, defining communication protocols and

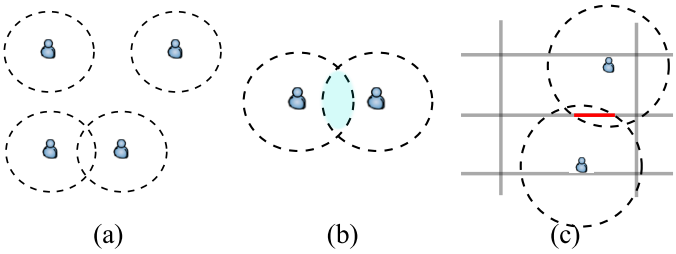


Fig. 12. Coverage metrics in the TOMP framework [53]: (a) the *Static Coverage* does not take into account any future movement, so nodes are considered in contact or not based on their present position; (b) the *Free Space Coverage* considers the possible movement of nodes in free space: the future meeting probability is the area of intersection of the two circles that represent the possible movements of the two nodes; (c) the *Graph-based Coverage* takes into account the underlying structure of road graph to limit the prediction to the road graph.

network architectures. In the following paragraphs, we will detail better the two approaches.

1) ***Subset Selection****:* Ioannidis et al. propose pushing updates of dynamic content from the infrastructure to end-users [48]. They assume that the cellular infrastructure has a fixed aggregate bandwidth that needs to be allocated between end-users. Peers exchange opportunistically any stored content between them. A rate allocation optimization is proposed to maximize the average freshness of content among all end-users. Two centralized and distributed algorithms are presented. Similarly, Han et al. and Li et al. tackle the offloading problem employing a subset selection mechanism based on the user contact pattern [49], [50]. While in the first work Han et al. study how to choose a subset of dimension k to be initially infected [49], Li et al. consider the optimal subset selection as an utility maximization problem under multiple linear constraints such as traffic heterogeneity, user mobility, and available storage [50]. The subset selection problem is NP-hard, similarly to the case of the minimum AP set-selection problem presented in [35] and discussed in Section IV-A. Both works propose *greedy* selection algorithms to identify a sub-optimal target set. A point in common for all the subset selection strategies is that the network provider should be able to collect information about node contact rates in order to compute the best subset.

Using social networking arguments, Barbera et al. analyze the contact pattern between end-users, in order to select a subset of central VIP users that are important for the network in terms of *centrality* and *page-rank* [51]. The key idea is to transform these few central VIP users into data forwarders between standard nodes and the Internet. The authors exploit the repetitive and periodic mobility of humans to train the selection algorithm to build the networks' social graph over which the VIPs selection is made. An analogous approach is exploited by Chuang et al., which merge the subset selection problem with the concept of social relationship between end-users [52]. They propose a community-based algorithm that selects users belonging to disjoint social communities as initial seeds, in order to maximize the offloading efficiency. In effect, the selection of initial seeds based only on encounter probability proves to be insufficient, as users with high encounter probability might belong to the same community. The goal is to select the set of initial sources so that both cellular traffic load and delivery time are minimized. Also in this schema, mobile end-users are required to upload periodic information on the most frequent contacts, in order to let the centralized algorithm to choose the best subset of seed users.

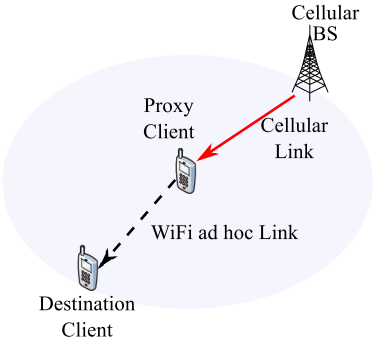


Fig. 13. Network extension in [62]: destination client experiences bad cellular connectivity . After the discovery of a neighbor node with better channel conditions, data is routed through this "proxy client" in the cellular network.

Baier et al. approach the subset selection problem by predicting the movement of end-users in order to estimate future inter-device connectivity [53]. The system, named TOMP (*Traffic Offloading with Movement Predictions*), retrieves information about actual positioning and speed of mobile devices rather than connectivity patterns. The framework selects as seed users the nodes that have the best future connectivity likelihood with other nodes based on movement prediction. As explained in Fig. 12, TOMP proposes three coverage metrics to predict the future movements of nodes: static coverage, free-space coverage, and graph-based coverage.

Discussion. Selecting high potential nodes as seeds of the dissemination process influences the performance of the offloading strategy. Wisely chosen seed users may infect a larger number of nodes, resulting in lesser late retransmissions. Subset selection algorithms commonly employ information on social interactions among users and their mobility patterns to figure out which nodes have the best features. Note that a control channel, binding the end-nodes to a central entity, is usually required in order to transfer context information. The performance of the offloading algorithm relies heavily on the understanding of the system dynamics. For this reason, it is essential to analyze how nodes meet creating communication opportunities in a fine-grained fashion, and characterize mobility at the microscopic level. Offloaded data vary from 30 to 50% for all the surveyed papers depending on the delay-tolerance and the dataset considered. However, apart one notable exception [52], only small scale and very specific datasets have been evaluated (typically around 100 users), providing a limited confidence in the generality of results.

2) ***Offloading Mechanisms****:* Luo et al. designed a new unified architecture for cellular and ad hoc networks, to leverage the advantages of each technology [62]. In this case, the goal is to increase the throughput experienced by mobile users by taking advantage of neighbors with better cellular connectivity, employed as a proxy. The working schema, as shown in Fig. 13, allows mobile users experiencing a low cellular downlink channel rate, to connect via ad hoc links

to a neighbor with better cellular channel conditions. The proxy node then acts as a gateway for data traffic of its peers. Data is further relayed through IP tunneling via intermediate relay clients to the destination, using the ad hoc link. The paper proposes also two proxy discovery protocols (namely *on demand* and *greedy*), and analyzes the impact of the proxy relaying schema on the cellular scheduling.

Mayer et al. propose a routing scheme for the offloading of unicast message exchange between end-users [63]. The offloading schema is based on a simple assumption: the higher the probability that a message can be delivered through the infrastructure in case of failing opportunistic delivery, the longer DTN routing takes to deliver the message. In effect, the protocol initially attempts to deliver messages through opportunistic communications and switches to the infrastructure network only when the probability of delivering the message within the deadline becomes unlikely. This opportunistic/infrastructure routing decision is taken locally exploiting information exchanged with other nodes upon encounters. Key contextual information includes awareness for destination node and infrastructure capabilities. In this way, the system tries to offer a reliable message delivery, while saving cellular traffic at the same time.

The Push-and-Track framework tackles the problem of disseminating popular content with guaranteed delays [64]. Fig. 14 presents the basic approach. A subset of users is initially infected through the infrastructure. The content is forwarded through T2T links when nodes meet. Mobile nodes, upon content reception, send a lightweight acknowledgment message to the coordinator through the cellular infrastructure. The central coordinator may re-inject copies if the diffusion status is low, in order to encourage the dissemination process. Acknowledgment messages may also contain information on encountered nodes and even the geographic position of the encounter. The monitoring mechanism allows the coordinator to have an up to date picture of the content dissemination status and to predict which nodes are the best to re-inject additional copies of the content, e.g., which uninfected nodes have the best potential to "boost" diffusion. Note that, in this case, the opportunistic dissemination decision is left to mobile users, and the coordinator only checks from time to time the diffusion status, possibly intervening by triggering re-injections. For instance, when the time gets closer to the delivery deadline, the coordinator enters a "panic zone", and content is pushed to all uninfected nodes through the infrastructure, in order to meet the delivery delay constraint. Since acknowledgment messages are much smaller than the actual content, the system allows significant reduction of the infrastructure load. Using the same framework, Rebecchi et al. propose a derivative-based re-injection strategy that exploits some characteristic properties of opportunistic data diffusion to optimize delivery [65]. Another approach in a similar line exploits a learning framework to understand when and to how many seeds content should be injected [66]. Interestingly, this class of offloading methods advertises high offloading efficiency (more than 50% of cellular traffic saved) even for very tight delivery delays.

Izumikawa et al. offer an offloading solution (called RoC-Net) that exploits the difference of traffic load among different locations [67]. Consider the distinct instantaneous traffic volume in a business district and a residential district during daytime. In case of localized RAN congestion, each delay-tolerant data request originated in that area, instead of being transmitted to the overloaded cellular BS, is forwarded to a neighbor that is likely to head toward a less congested area, as shown in Fig. 15. A particle filter is employed to predict future movement pattern of neighbor users, starting from its movement history. When a terminal is in a congested area, a coefficient of variation is exchanged upon opportunistic meeting with neighbors, to decide which user is more likely to move to a low-congested area.

Finally, some architectures exploit the availability of hybrid delivery options (Cellular, APs, and opportunistic). Pitkanen et al. describe a system to extend the range of fixed WiFi APs through the DTN approach [68]. Delay-tolerant data is shifted from the cellular network to the closest WiFi AP, contributing to preserve cellular bandwidth for real-time and interactive applications. Similarly, Petz et al. introduce MAD-Server, an offloading-aware server that enables the distribution of web-based content through a multitude of access networks

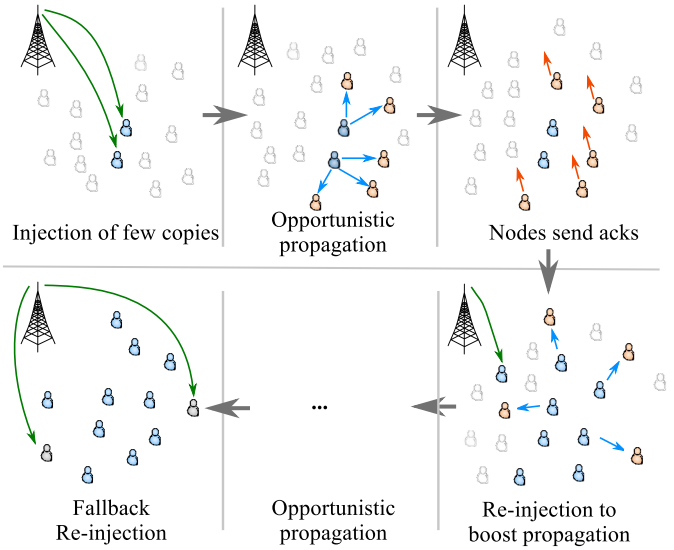


Fig. 14. Push-and-Track framework [64]. A subset of users receive the content from the infrastructure channel and start diffusing it opportunistically. Nodes acknowledge content reception to the source, allowing it to keep track of the content dissemination status. The source may also re-inject copies if the diffusion status is low, in order to feed the dissemination process. Finally, the content is pushed to uninfected nodes.

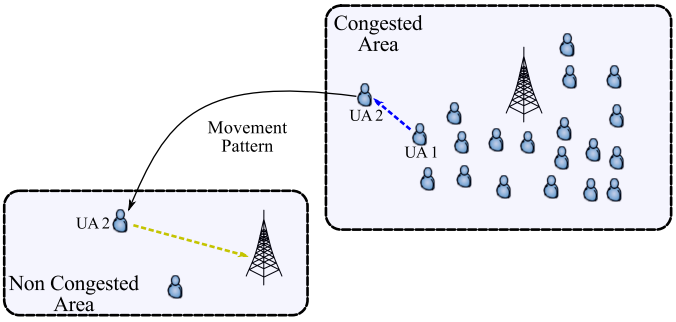


Fig. 15. High level overview of RocNet. UA 1 is in a congested area. Upon discovering, UA 1 forwards data to UA 2 if it is more likely to move to a non congested area than UA 1.

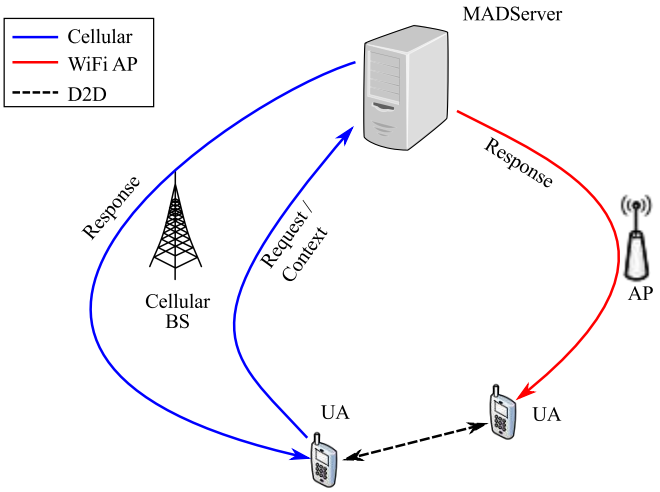


Fig. 16. MADServer architecture [69]. Requests are always routed through the cellular channel, along with contextual information. Responses may transit on cellular, WiFi or opportunistic channels, depending on their size and delay-tolerance.

(Fig. 16) [69]. Both systems use contextual information from users, to predict where to cache data in advance, and are able to split the content into multiple pieces, independently delivered on different access networks. Small and time critical content is always transmitted over the cellular infrastructure, while large data, such as videos and pictures are offloaded only when it is beneficial and within deadline.

Discussion. The definition of network architectures capable of exploiting different technologies to deliver content is a key milestone for the research community. The current trend is toward network-aided offloading schemes, where the cellular network guides its connected peers in the neighbor discovery and connectivity management phase. The routing scheme takes advantage of well-placed neighbors used as preferred gateways for data forwarding. The substantial use of context information harvested from end-users, or exchanged locally, is exploited to drive the routing decision through the optimal interface. Future challenges include the development of novel coordination mechanisms and inter-technology scheduling policies to control content retrieval between multiple access technologies and opportunistic networks. Cellular operators are particularly interested in the development of innovative capacity models able to predict the additional gains provided by the activation of offloading, and to plan how much traffic they can divert from their core network.

V. Target Offloading Architecture

The analysis conducted so far reveals that the various forms of offloading are quite different, in terms of both network infrastructures and delivery delay requirements. Despite this, it is still possible to identify from the specific solutions a number of common functionalities making up an advanced offloading scheme. The challenge is to go beyond what is done today, which is mainly a user-initiated offloading process. The opportunity for operators to drive the offloading process will provide them with better network management options.

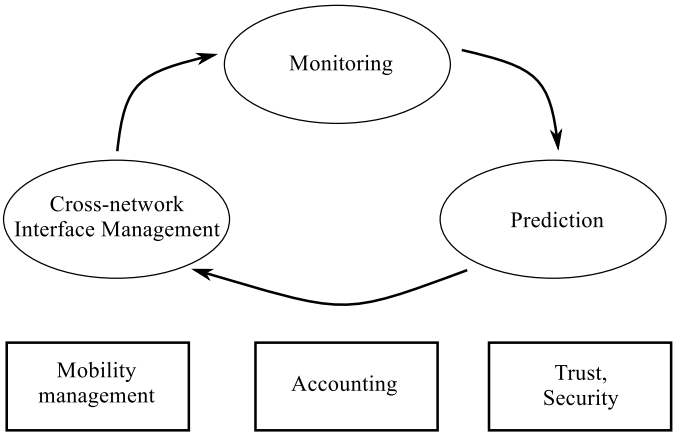


Fig. 17. Offloading coordinator functional building blocks.

A. Functional Architecture

In order to make this vision possible, we need to extract a number of generic high-level functionalities that make up the offloading system. This analysis is significant in view of the integration of offloading capabilities into future mobile networks architecture. Fig. 17 provides a high-level scheme to help us drive the discussion. Most of the works we surveyed consider an *offloading coordinator*, an entity specifically dedicated to the implementation of the actual offloading strategy. Its main task is to pilot the offloading operation depending on network conditions, users' requests, and operator offloading policy. While conceptually represented by a single entity, its physical location in the network may vary, and sometimes its implementation could be totally distributed. However, it is possible to identify, among all, three main interdependent functional blocks for the offloading coordinator: (i) monitoring, (ii) prediction, and (iii) cross-network interface management.

Monitoring. Monitoring provides methods to track the actual data propagation spreading, user's requests, and to retrieve contextual information from nodes and the network. Retrieved information is necessary to evaluate and execute the offloading strategy. The monitoring block often requires the presence of a persistent control channel that allows end-users to interact with the offloading coordinator (e.g., the cellular channel is explicitly employed with this purpose in [5], [64]). Harvested information is then passed to the prediction block to be processed.

Prediction. Prediction relates to the ability of the offloading coordinator to forecast how the network will evolve based on past observations. Typical prediction deals with mobility [20], [53], contact patterns of users [5], or expected throughput [47], [25]. Such predictions are then used to pilot the entire offloading process more efficiently. This is the block where typically the offloading intelligence resides. The complexity of the prediction should trade off its applicability, in order to guarantee the real-time operation of the offloading process. Predicted values are transmitted to the interface management block in order to drive the offloading process.

Cross-network interface management. Traditional approaches manage each interface independently. However, integrated management allows exploiting in parallel the benefit of

each available interface. Cross-network interface management deals with deciding on which network the required content (or parts of it) will flow. Concepts such as load balancing, throughput maximization, congestion control, and user QoE (Quality of Experience) relates to this functional block. By exploiting this information, the network itself will be able to identify the current situation and optimize its performance. For instance, ANDSF and IFOM already use this capability [29], [30]. They are able to shift selected data on a given network interface, in order to obtain a benefit.

Additional transversal subjects emerge from the analysis of the literature. For instance, *mobility management*, *accounting*, and aspects related to *trust and security* are essential to support offloading strategies in mobile network architectures. Mobility management involves the seamless handover between different base stations due to the mobility of users. Accounting functionalities enable proper accounting and charging information for the offloaded traffic and users. This is a key component in order to design incentive mechanisms to stimulate the participation of mobile users in the offloading process. Finally, trust and security mechanisms guarantee the privacy and the integrity of both infrastructure and D2D communications. This block is essential since most offloading strategies transform the user into an active network element.

These can be regarded as the basic functional building blocks that mobile networks should provide to ensure offloading capabilities. Anyway, we stress that, depending on the specific implementation, the proposed functionalities may be present or not. For instance, mobility management modules are elemental in non-delayed offloading, in order to handle the handover between different APs, and to secure continuity of ongoing data session. On the other hand, the same block could be disregarded when dealing with delayed D2D transmissions.

VI. Assessing Mobile Data Offloading: Metrics and Evaluation Tools

It is quite challenging to compare the performance of different offloading strategies only on the basis of the results reported in the literature, because the evaluated metrics often differ. In addition, we can assess the performance of offloading from the perspectives of both network operators and users, which have essentially divergent needs [11]. In this section, we will give hints on the metrics that we believe important for the evaluation of offloading strategies. In addition, we discuss simulators, mobility models, and testbeds, which play a significant role in performance evaluation.

A. Metrics

From a cellular operator's point of view, offloading should serve as a reserve of capacity, which may be added to the network in case of heavy congestion. For this reason, a significant challenge is to quantify the additional capacity brought by the use of offloading strategies. The most notable effect of offloading should be the reduction of traffic load and congestion in the primary network. Nevertheless, capacity improvements depend, among other things, on the number of mobile devices or wireless APs involved in the process, on the mobility of nodes, on the size and the delay-tolerance of the offloaded content. On the other hand, user satisfaction is often associated with Quality of Experience (QoE), so the received throughput and timely reception parameters are regarded as the most important parameters. Commonly employed metrics of interest today in the literature are the following:

Offloading Ratio or **Offloading Efficiency.** It is the fundamental parameter to evaluate the effectiveness of any offloading strategy from an operator point of view. It is measured as the ratio of the total traffic offloaded (transferred through alternative channels) to the total traffic generated [12], or as the ratio of the total load of traffic that flows on the cellular channel after the offloading process to the traffic on the infrastructure in the absence of any offloading strategy [64].

Offloading Overhead. The offloading overhead metric evaluates, in a broad sense, the amount of additional control data required by the offloading mechanism. For instance, as explained by Sankaran [90], in the IFOM scenario, the overhead is represented by the messages needed to exchange and discover IFOM capabilities between involved nodes. In the Push-and-Track scenario, the offloading overhead depends on the control traffic that flows into the infrastructure channel, intended to pilot the offloading process [64].

Quality of Experience (QoE). From a user perspective, the most critical metric is the Quality of Experience (QoE), which is linked to its satisfaction. For any offloading class, the total achievable throughput is a common but important metric. The QoE indicator is then made up of several sub-metrics that depend on the application and the type of offloading. For instance, video streaming QoE-metrics are the Peak Signal-to-Noise Ratio (PSNR) and the amount of packet loss. In delayed offloading, the delivery time is the most meaningful metric, representing the amount of time before content reception.

Power Savings. In some works, the concept of offloading is associated with the power savings that may be attained by the nodes. This is possible because the WiFi interface is more efficient in terms of energy per bit than the cellular interface. Traffic offloading algorithms are interesting to achieve energy savings.

Fairness. Fairness in terms of resource usage (in particular energy consumption) can be an important evaluation parameter. Fairer systems tend to distribute resources uniformly without relying too much on the same users. This aspect is critical in D2D offloading, where an unbalanced use of resources could lead to premature battery depletion. For instance, seed-based offloading strategies risk being unfair, because data is transmitted to a limited number of users that retransmit it on the secondary channel. Even if this strategy could reduce the overall energy consumption, it is unfair in terms of user's individual energy consumption.

As a summary, we subdivide the surveyed papers with their evaluated metrics in Table III. In the last column, we include how the performance evaluation has been executed: simulation, real testbed, or analytical study. Regarding simulation studies, some works do not explicitly specify which tool has been

used. It is also important to note that evaluated metrics often depends on how performance is assessed. In particular, power saving is commonly evaluated in experimental works, while offloading efficiency is typically estimated through simulation. In general, simulation-based evaluations are likely to propose a system-wide approach, i.e., they consider the whole network, even with some approximation. On the other hand, evaluations based on real experiments, due to the inherent complexity of assembling large-scale scenarios, focus more on terminal-level parameters and small-scale experiments.

B. Simulation Tools

Simulations play a major role in analyzing performance of offloading strategies and protocols. We observe from the last column of Table III that no simulator is predominant. Four classes of simulators emerge: MATLAB, ns-2 [130], ONE [131], and custom-made (mainly Java and C-based). MATLAB is usually employed for the evaluation of radio signal propagation and queue-based models. Ns-2 is a mature open-source network simulator, and serves as a generic platform for packet-level analysis. Surprisingly, none of the work that we surveyed makes use of ns-3 [132], an evolution of ns-2. The ONE is essentially an opportunistic network simulator for DTNs. It already implements some typical DTN routing protocols and mobility models.

Despite the availability of these simulators, which are tested, their complexity, a limited support to unusual mode of operation (as in the case of offloading) and the inability to evaluate large-scale deployment has meant that many works rely on custom simulators, mainly written in Java and C. This could result in a problem for reproducibility of results and for the construction of a common software base in the future.

C. Mobility Models

Mobility of nodes is at the base of performance of offloading strategies. Mobility models can be extracted from mathematical random process, such as *random waypoint* (RWP) or *random walk* (RW). While simple to implement, these models are not realistic in reproducing human behavior [133]. Map based mobility model (MBM), shortest path map-based model (SPMBM), route-based model (RBMs), or movements based on human activities including work day movement (WDM) [134], try to improve realism exploiting information from real-world behavior of humans. Nevertheless, finding realistic mobility models is a complex challenge. For this reason, the analysis is often made taken real world traces from CRAWDAD [135], a platform to share wireless network traces. Unfortunately, available traces often present a limited number of users (less than 100 in most of the case), and have a low spatial and temporal granularity.

D. Testbeds Implementation

Despite the fact that there is a huge body of literature devoted to offloading, most of the evaluations are simulation-based. Although simulation permits to evaluate many different aspects, its effectiveness is intrinsically limited by the simplifications of the model and the software complexity, to allow processing in a reasonable time.

On the other hand, real-world experimentation permits to complement the simulative and numerical analysis, enabling to evaluate also the impact of complex phenomena that happens in the wireless medium, such as interference, scheduling and overhead, often inaccurately modeled in simulation. Numerous technical challenges arise in real-world testbed implementation. For instance, there is currently no mobile system able to accommodate the cross-layer requirements of offloading. Thus, implemented testbeds have different constraints depending on the design choices that pose specific limitations on system performance. In most of the case, the testbed is developed employing standard designs for the PHY and MAC layers, and modifying the above layers. This narrows down the possible degrees of freedom, because only part of the networking stack is accessible.

As an example, Android, which is one of the most open environments to develop mobile applications, does not expose APIs to switch the IEEE 802.11 interface into the ad hoc mode. First, the device should be *rooted* to gain administrative access rights. A specific *Linux wireless tools* package has to be specifically compiled for ARM-based devices and installed into the device. Then, to connect the device to an ad hoc network, the Android NDK (Native Development Kit) has to be employed [136].

VII. Cellular Networks and the Bandwidth Crunch: Other Possible Solutions

As already mentioned, the intricate problem of mobile data explosion can be addressed in several ways. Hence, we briefly review alternative solutions to the capacity problem in cellular networks linked to data offloading. We identified five main categories related to data offloading, each one bringing advantages and disadvantages:

- Addition of small-size base station and/or femtocells.
- Multicasting/broadcasting data inside the cell.
- Integration of cognitive radio mechanisms.
- Proactive pushing of popular content on devices.

It is worth to note that many of these possibilities are orthogonal to each other, and can be deployed at the same time. In addition, the methods outlined in this section may also complement the strategies presented along the survey.

A. Small-Sized and Femto-cell Deployment

The first solution adopted by the majority of cellular providers to face data growth is to scale the RAN by building more base stations with smaller cell size. Reducing the size of macro-cell increases the available bandwidth and cuts down the transmission power [137]. An obvious drawback is that operators have to build additional base stations. Equipment costs, site rental, backhaul, and power consumption, make this strategy very expensive in terms of CAPEX and OPEX. In addition, according to [138], only a small fraction of mobile users (around 3%) consume more than 40% of all mobile traffic. Consequently, the majority of users gets only a minimal

TABLE III: Summary of key mobile data offloading strategies.

Ref.	Strategy	Delay Requirements	Evaluated Metrics	Performance assessment
[12] Lee et al.	AP-based	Non-delayed	Efficiency	Simulation (MATLAB)
[13] Fuxjager et al.	AP-based	Non-delayed	Efficiency	Testbed
[14] Liu et al.	AP-based	Non-delayed	Efficiency, QoE (Availability)	Testbed
[15] Hu et al.	AP-based	Non-delayed	QoE (SINR,throughput), Fairness	Simulation (MATLAB)
[16] Ristanovic et al.	AP-based	Non-delayed, delayed	Efficiency	Simulation (Java)
[17] Bulut et al.	AP-based	Non-delayed	Efficiency	Simulation
[18] Mehmeti et al.	AP-based	Non-delayed	Efficiency, QoE (completion time)	Analytical model
[19] Singh et al.	AP-based	Non-delayed	QoE (SINR)	Simulation, Analytical model
[28] Hagos et al.	AP-based	Non-delayed	QoE (SINR), Efficiency	Simulation (MATLAB)
[32] Makaya et al.	AP-based	Non-delayed	QoE (throughput), Power Saving	Testbed
[40] Hou et al.	AP-based	Non-delayed	QoE (throughput)	Testbed
[41] Patino Gonzalez	AP-based	Non-delayed	QoE (throughput), Power Saving	None
[42] Nirjon et al.	AP-based	Non-delayed	QoE (throughput), Power Saving, Overhead	Testbed
[43] Rahmati et al.	AP-based	Non-delayed	QoE (throughput)	Testbed, Simulation
[44] Kang et al.	D2D	Non-delayed	Cost	Simulation
[45] Zhu et al.	D2D	Non-delayed	Power Saving	Testbed, Simulation (ns-2)
[46] Leung et al.	D2D	Non-delayed	QoE (completion time, throughput), Cost, Fairness	Simulation (C++)
[47] Stiemerling et al.	D2D	Non-delayed	QoE (packet loss, completion time)	Simulation (C++, Java)
[54] Karunakaran et al.	D2D	Non-delayed	Power Saving, QoE (Throughput)	Simulation, Analytical model
[55] Hua et al.	D2D	Non-delayed	QoE (PSNR)	Simulation (OPNET)
[56] Seferoglu et al.	D2D	Non-delayed	QoE (Throughput)	Analytical model, Testbed
[57] Keller et al.	D2D	Non-delayed	QoE (Throughput), Power Saving	Testbed
[58] Ramadan et al.	D2D	Non-delayed	Power Saving	Testbed
[59] Sharafeddine et al.	D2D	Non-delayed	Power Saving	Testbed
[60] Andreev et al.	D2D	Non-delayed	QoE (Throughput), Power saving	Simulation, Testbed
[70] Doppler et al.	D2D	Non-delayed	QoE (Throughput)	Simulation
[74] Zulhasnine et al.	D2D	Non-delayed	QoE(Throughput)	Simulation (C++)
[75] Yu et al.	D2D	Non-delayed	QoE (Throughput), Power saving	Analytical model, Simulation
[76] Malandrino et al.	D2D	Non-delayed	Efficiency	Simulation
[77] Hasan et al.	D2D	Non-delayed	QoE (Throughput)	Simulation (Matlab)
[78] Li et al.	D2D	Non-delayed	Efficiency, QoE (Throughput)	Simulation
[79] Li et al.	D2D	Non-delayed	Efficiency, Fairness	Simulation
[80] Yaacoub et al.	D2D	Non-delayed	Power Saving, Fairness	Simulation (Matlab)
[20] Siris et al.	AP-based	Delayed	Efficiency, QoE (completion time)	Simulation
[21] Chen et al.	AP-based	Delayed	Efficiency, QoE (throughput)	Testbed
[5] Dimatteo et al.	AP-based	Delayed	Efficiency, QoE (user satisfaction)	Simulation
[22] Ra et al.	AP-based	Delayed	Power Saving, QoE (completion time)	Simulation, Testbed
[23] Go et al.	AP-based	Delayed	Efficiency, QoE (completion time)	Simulation
[25] Balasubramanian et al.	AP-based	Delayed	Efficiency, QoE (Completion time)	Simulation
[26] Yetim et al.	AP-based	Delayed	Efficiency	Simulation
[34] Mehmeti et al.	AP-based	Delayed	Efficiency, QoE (completion time)	Analytical model
[35] Trestian et al.	AP-based	Delayed	Efficiency, QoE (completion time)	Simulation
[39] Malandrino et al.	AP-based	Delayed	Efficiency	Simulation
[37] Abdrabou et al.	AP-based	Delayed	QoE (completion time)	Simulation (ns-2)
[38] Malandrino et al.	AP-based	Delayed	Efficiency, QoE (throughput)	Simulation
[48] Ioannidis et al.	D2D	Delayed	Efficiency, QoE (completion time)	Analytical model
[49] Han et al.	D2D	Delayed	Efficiency, Power Saving	Simulation (C), Testbed
[50] Li et al.	D2D	Delayed	Efficiency	Simulation
[51] Barbera et al.	D2D	Delayed	Efficiency	Simulation
[52] Chuang et al.	D2D	Delayed	Efficiency, QoE (completion time)	Simulation
[53] Baier et al.	D2D	Delayed	Efficiency, QoE (throughput)	Simulation (ns-2)
[62] Luo et al.	D2D	Delayed	QoE (throughput), Overhead	Simulation (ns-2)
[129] Busanelli et al.	D2D	Delayed	QoE (completion time), Overhead	Testbed
[63] Mayer et al.	D2D	Delayed	Efficiency, QoE (completion time)	Simulation (ONE)
[64] Whitbeck et al.	D2D	Delayed	Efficiency	Simulation (Java)
[65] Rebecchi et al.	D2D	Delayed	Efficiency, Overhead	Simulation (Java)
[66] Valerio et al.	D2D	Delayed	Efficiency	Simulation
[67] Izumikawa et al.	D2D	Delayed	Efficiency, Fairness	Simulation (ONE)
[69] Petz et al.	D2D	Delayed	Efficiency, QoE (completion time)	Testbed

benefit from this strategy, as heavy consumers will continue to grasp the bulk of the bandwidth.

Another possibility is to push the adoption of femtocells. The approach is analogous to AP-based offloading but makes use of the same access technology of the macro-cell. However, since femtocells work on the same frequency as the macro network, interference management becomes challenging [139]. Performance of femtocell-oriented offloading is investigated in [140], [141]; other works compare the gains brought by femtocells against AP-based offloading [15], [142]. Energy-related topics are presented in [143]. Interested readers should also refer to existing surveys on femtocells in the literature [144], [145]. The trend toward smaller cells is part of the so-called *HetNet* paradigm, in which cellular macro-cells coexist and overlay a myriad of smaller cells. This affects the design of the resource allocation scheme, and ongoing researches focus on the decision if a user should be served by the macro or by a closer small-cell. A flexible small-cell deployment helps in eliminating coverage holes, and increasing the network capacity in some regions inside a macro-cell [146].

B. Multicast/Broadcast

When many users in spatial proximity ask for the same data, multicast could emerge as a good alternative to data offloading for comparable use cases. Multicast employs a single radio link, shared among several users within the same radio cell.[8] Logically there is no interaction, and users can only receive content. Multicast is a clever strategy to provide content to multiple users exploiting redundancy of requests, allowing in principle great resources saving.

Besides requiring modifications in the cellular architecture, multicast has intrinsic and still unresolved inefficiencies that limit its exploitation. Each user experiences different radio link conditions. This variability heavily reduces the effectiveness of multicast, since the base station must use a conservative modulation to ensure a successful to each user. Nodes that are closer to the base station are able to decode data at a higher rate, while others located near the edge of the cell have to reduce their data rate. Thus, the worst channel user dictates the performance, lowering the overall multicast throughput. This is the main reason why offloading can be beneficial also in case of multicast, as demonstrated in [115].

C. Cognitive Radio Integration

The spectrum of frequencies available to mobile operators is already overcrowded, while other portions of the spectrum are relatively unused. The limited available bandwidth and the inefficiency in its use call for an opportunistic use of unoccupied frequencies [147]. Cognitive radios could dynamically detect unused spectrum and share it without harmful interference to other users, to shift data on it, enhancing the overall network capacity [148]. Cognitive radio can be employed to offload cellular networks [149], in cohabitation with the HetNet paradigm [150]. Cognitive technologies are thus capable of increasing spectrum efficiency and network capacity significantly.

[8]LTE proposes an optimized broadcast/multicast service through *enhanced Multimedia Broadcast/Multimedia Service* (eMBMS) [114].

D. Proactive Caching

Caching is a popular technique, commonly employed in web-based services in order to reduce traffic volume, the perceived delay, and the load on servers. Caching techniques work by storing popular data in a cache located at the edge of network. Some of these classical concepts can be re-utilized in mobile networks to tackle congestion at RAN. In order to avoid peak traffic load and limit congestion in mobile networks, techniques for predicting users' next requests and pre-fetching the corresponding content are available [151], [152], [153]. Data may be pro-actively cached directly at the user device, at cellular base station, or at IEEE 802.11 APs to improve the offloading process. The prediction is performed using statistical methods or machine learning techniques, and its accuracy is a key factor in performance. Note that some of these techniques may be (or are already) used in the delayed AP-based offloading schemes considered in Section IV-A.

VIII. Open Challenges

Mobile data offloading remains a new and very hot topic, frequently identified as one of the enablers of next-generation mobile networks. Future research directions are manifold. Effective offloading systems require a tighter integration within the 3GPP and the wireless broadband infrastructures. Additional features still need to be developed to handle mobility of users, distributed trust, session continuity, and optimized scheduling policies. Offloading strategies may take advantage both of the AP connectivity and terminal-to-terminal communication opportunities. A very interesting future research area concerns how to merge, in a fully integrated network architecture, the different and often stand-alone offloading possibilities presented along this survey. This unified architecture requires reconsidering existing wireless network paradigms. Therefore, future cellular architectures should intelligently support the distribution of heterogeneous classes of services, including real-time and delay-tolerant flows, to cope with an overall traffic increase of several orders of magnitude. If we consider delayed offloading, there is not yet a clear consensus of how network operators can drive the offloading process, assisting users in the opportunistic data retrieval, guaranteeing satisfaction, and maximizing at the same time the amount of saved traffic. A fine comprehension of data traffic and mobility patterns of nodes is required. It is critical to understand which types of traffic can be safely diverted on complementary data channels and which cannot, based on their delivery requirements. Additionally, fundamental research should focus on how nodes move and meet, creating communication opportunities in a fine-grained fashion.

Besides research challenges, the implementation of offloading strategies results in a variety of practical challenges. Academia and industry must tackle such challenges in order to make offloading a viable answer to the mobile data overload problem. To date, both technical and adoption-related challenges complicate the widespread introduction of offloading.

The foremost technical challenge is related to the lack of a widely accepted mechanism to handle transparently several flows in parallel on different interfaces (nor protocols resilient to link failures, communication disruptions, and capable of handling substantial reception delays). As pointed out throughout the survey, various mechanisms have been proposed, but there is not yet a consensus on a de-facto standard. From the user perspective, a major concern comes from the dramatic battery drain of multiple wireless interfaces simultaneously turned on, even in idle mode. As of today, this combined use will seriously reduce the battery life of mobile devices. Possible solutions may be the design of low-powered network interfaces or the implementation of energy saving policies (a sort of duty cycle to switch on and off network interfaces), although privacy concerns prevent network operators to force a device to turn on and off a network interface.

Regarding user adoption, we should not forget that user collaboration, especially in the opportunistic approach, is essential for any offloading strategy. In order to make offloading feasible, end-users must accept to share some resources (battery, storage space, etc.), and their wireless interface should be turned on. The central question here is how to motivate users to participate. Mobile operators should propose a business concept for rewarding their customers, to make offloading attractive and fully functional at the same time with user participation. A sufficient number of game theory-based works attempt to clarify the relationship between the proposed incentives and the expected offloading benefit [27], [154], [155]. Additional issues lie on the security and privacy plan of users employing mobile-to-mobile transmissions. Users rarely accept anyone stranger to access data stored on their devices. Further challenges include the development of an infrastructure to ensure distributed trust and security to terminals involved in the offloading process.

A key question in mobile data offloading concerns the role of the network provider in the offloading process. Should the operator drive carefully the offloading process, or are end-nodes sufficiently autonomous to decide for themselves the best offloading strategy? In other words, future implementations should clarify how much the offloading process will be user-driven or operator-driven. Both strategies present advantages and disadvantages. An operator may have a better view of the overall network, while a user may have only a local and obviously partial view. On the other hand, an operator-driven offloading strategy may tend to give priority to a certain type of traffic or class of users, while a distributed offloading strategy may result more fair. The debate on these issues is still very open, and more research is needed along the lines introduced in this survey.

IX. Conclusion

Mobile data offloading has the potential to relieve the cellular network congestion at minimal cost, allowing users to experience high quality network access and contributing to solve the longstanding RAN overloading problem. The discussion provided in this survey strongly advocates the use of alternative mobile access networks for offloading purposes. We investigated the concept of mobile data offloading, identifying its key benefits, technological challenges, and current research directions. In particular, after presenting a broad classification of current offloading strategies based on their requirements in terms of delivery guarantee, we presented the technical aspects and the state of the art for two main approaches. The former is more mature and proposes a tight integration between the cellular RAN and a complementary access network, allowing for real-time data offloading. The latter, still experimental, exploits the delay tolerance of some types of data to optimize their delivery. We identified some common functional blocks, proposing a general high-level architecture valid for any mobile data offloading system. We further investigated open research and implementation challenges and the existing alternatives to mitigate the cellular overload problem.

Acknowledgment

Filippo Rebecchi and Marcelo Dias de Amorim carried out part of the work at LINCS (http://www.lincs.fr). This work is partially supported by the European Commission in the framework of the FP7 Mobile Opportunistic Traffic Offloading (MOTO 317959), by the EINS (FP7-FIRE 288021), and EIT ICT Labs MOSES (Business Plan 2014) projects.

References

[1] Cisco, "Cisco visual networking index: Global mobile data traffic forecast update (2013 – 2018)," 2014.

[2] P. Taylor, "Data overload threatens mobile networks," accessed: 2013-08-21. [Online]. Available: http://www.ft.com/intl/cms/s/0/caeb0766-9635-11e1-a6a0-00144feab49a.html

[3] B. G. Mölleryd, J. Markendahl, J. Werding, and O. Mäkitalo, "Decoupling of revenues and traffic - is there a revenue gap for mobile broadband ?" in *Conference on Telecommunications Internet and Media Techno Economics (CTTE)*, Ghent, Belgium, Jun. 2010, pp. 1–7.

[4] S. Curtis, "Can you survive on 4g alone?" accessed: 2013-11-06. [Online]. Available: http://www.telegraph.co.uk/technology/internet/10272292/Can-you-survive-on-4G-alone.html

[5] S. Dimatteo, P. Hui, B. Han, and V. O. K. Li, "Cellular traffic offloading through WiFi networks," in *IEEE International Conference on Mobile Ad-hoc and Sensor Systems (MASS)*, Valencia, Spain, Oct. 2011.

[6] L. Korowajczuk, *LTE, WiMAX and WLAN Network Design, Optimization and Performance Analysis*. John Wiley & Sons, 2011.

[7] "Data Offload – Connecting Intelligently," White Paper, Juniper Research, 2013.

[8] iPass, "iPass application." [Online]. Available: http://www.ipass.com/

[9] Guglielmo, "BabelTen application," accessed: 2013-08-21. [Online]. Available: http://www.guglielmo.biz/Servizi.aspx?lan=eng

[10] W. Mohr and W. Konhauser, "Access network evolution beyond third generation mobile communications," *IEEE Communications Magazine*, vol. 38, no. 12, pp. 122–133, Dec. 2000.

[11] A. Aijaz, H. Aghvami, and M. Amani, "A survey on mobile data offloading: technical and business perspectives," *IEEE Wireless Communications*, vol. 20, no. April, pp. 104–112, 2013.

[12] K. Lee, J. Lee, Y. Yi, I. Rhee, and S. Chong, "Mobile data offloading: How much can WiFi deliver?" *IEEE/ACM Transactions on Networking*, vol. 21, no. 2, pp. 536–550, Apr. 2013.

[13] P. Fuxjager, I. Gojmerac, H. R. Fischer, and P. Reichl, "Measurement-based small-cell coverage analysis for urban macro-offload scenarios," in *Vehicular Technology Conference (VTC Spring)*, Yokohama, Japan, May 2011, pp. 1–5.

[14] S. Liu and A. Striegel, "Casting Doubts on the Viability of WiFi Offloading," in *ACM SIGCOMM workshop on Cellular networks: operations, challenges, and future design*, Helsinki, Finland, 2012, pp. 25–30.

[15] L. Hu, C. Coletti, N. Huan, I. Z. Kovács, B. Vejlgaard, R. Irmer, and N. Scully, "Realistic indoor wi-fi and femto deployment study as the offloading solution to lte macro networks," in *IEEE Vehicular Technology Conference (VTC Fall)*, Quebec City, QC, Sep. 2012, pp. 1–6.

[16] N. Ristanovic, J.-Y. Le Boudec, A. Chaintreau, and V. Erramilli, "Energy efficient offloading of 3G networks," in *IEEE International Conference on Mobile Ad-Hoc and Sensor Systems (MASS)*, Valencia, Spain, Oct. 2011, pp. 202–211.

[17] E. Bulut and B. K. Szymanski, "Wifi access point deployment for efficient mobile data offloading," in *ACM international workshop on Practical issues and applications in next generation wireless networks*, Istanbul, Turkey, 2012, pp. 45–50.

[18] F. Mehmeti and T. Spyropoulos, "Performance analysis of on-the-spot mobile data offloading," in *IEEE GLOBECOM*, Atlanta, GA, Dec 2013, pp. 1577–1583.

[19] S. Singh, H. Dhillon, and J. Andrews, "Offloading in heterogeneous networks: Modeling, analysis, and design insights," *IEEE Transactions on Wireless Communications*, vol. 12, no. 5, pp. 2484–2497, May 2013.

[20] V. Siris and D. Kalyvas, "Enhancing mobile data offloading with mobility prediction and prefetching," in *ACM international workshop on Mobility in the evolving internet architecture (MobiArch)*, Istanbul, Turkey, 2012, pp. 17–22.

[21] B. B. Chen and M. C. Chan, "Mobtorrent: A framework for mobile internet access from vehicles," in *IEEE INFOCOM*, 2009, pp. 1404–1412.

[22] M.-R. Ra, J. Paek, A. B. Sharma, R. Govindan, M. H. Krieger, and M. J. Neely, "Energy-delay tradeoffs in smartphone applications," in *International conference on Mobile systems, applications, and services - MobiSys*, San Francisco, CA, 2010, pp. 255 – 270.

[23] Y. Go, Y. G. Moon, and K. S. Park, "Enabling DTN-based data offloading in urban mobile network environments," in *International Conference on Future Internet Technologies*, Seoul, Korea, 2012, p. 48.

[24] F. Malandrino, C. Casetti, C. Chiasserini, and M. Fiore, "Offloading cellular networks through ITS content download," in *IEEE Conference on Sensor, Mesh and Ad Hoc Communications and Networks (SECON)*, Seoul, South Korea, Jun. 2012, pp. 263–271.

[25] A. Balasubramanian, R. Mahajan, and A. Venkataramani, "Augmenting mobile 3G using WiFi," in *ACM Mobisys*, San Francisco, CA, Jun. 2010.

[26] O. B. Yetim and M. Martonosi, "Adaptive usage of cellular and WiFi bandwidth: An optimal scheduling formulation," in *ACM CHANTS*, Istanbul, Turkey, Aug. 2012.

[27] D. Zhang and C. Yeo, "Optimal handing-back point in mobile data offloading," in *IEEE Vehicular Networking Conference (VNC)*, Nov. 2012, pp. 219–225.

[28] D. Hagos and R. Kapitza, "Study on performance-centric offload strategies for lte networks," in *6th Joint IFIP Wireless and Mobile Networking Conference (WMNC)*, Dubai, 2013, pp. 1–10.

[29] Y. M. Kwon, J. S. Kim, J. Gu, and M. Y. Chung, "ANDSF-based congestion control procedure in heterogeneous networks," in *IEEE International Conference on Information Networking (ICOIN)*, Bangkok, 2013, pp. 547–550.

[30] A. de la Oliva, C. Bernardos, M. Calderon, T. Melia, and J. Zuniga, "Ip flow mobility: Smart traffic offload for future wireless networks," *IEEE Communications Magazine*, vol. 49, no. 10, pp. 124–132, Oct. 2011.

[31] S. Frei, W. Fuhrmann, A. Rinkel, and B. V. Ghita, "Prospects for wlan in the evolved packet core environment," in *International Conference on New Technologies, Mobility and Security (NTMS)*, Istanbul, Turkey, May 2012, pp. 1–5.

[32] C. Makaya, S. Das, and F. J. Lin, "Seamless data offload and flow mobility in vehicular communications networks," in *IEEE Wireless Communications and Networking Conference Workshops (WCNCW)*, Paris, France, Apr. 2012, pp. 338–343.

[33] J. Korhonen, T. Savolainen, A. Ding, and M. Kojo, "Toward network controlled ip traffic offloading," *IEEE Communications Magazine*, no. 3, pp. 96–102, Mar. 2013.

[34] F. Mehmeti and T. Spyropoulos, "Is it worth to be patient? analysis and optimization of delayed mobile data offloading," in *IEEE INFOCOM*, Toronto, ON, Apr. 2014, pp. 2364 – 2372.

[35] I. Trestian, S. Ranjan, A. Kuzmanovic, and A. Nucci, "Taming the mobile data deluge with drop zones," *IEEE/ACM Transactions on Networking*, vol. 20, no. 4, pp. 1010–1023, Aug. 2012.

[36] C. Lochert, B. Scheuermann, C. Wewetzer, A. Luebke, and M. Mauve, "Data aggregation and roadside unit placement for a vanet traffic information system," in *ACM international workshop on VehiculAr Inter-NETworking*, San Francisco, CA, 2008, pp. 58–65.

[37] A. Abdrabou and W. Zhuang, "Probabilistic delay control and road side unit placement for vehicular ad hoc networks with disrupted connectivity," *IEEE Journal on Selected Areas in Communications*, vol. 29, no. 1, pp. 129–139, 2011.

[38] F. Malandrino, C. Casetti, C. Chiasserini, C, and M. Fiore, "Optimal content downloading in vehicular networks," *IEEE Transactions on Mobile Computing*, vol. 12, no. 7, pp. 1377–1391, 2013.

[39] F. Malandrino, C. Casetti, C. Chiasserini, C. Sommer, and F. Dressler, "Content downloading in vehicular networks: Bringing parked cars into the picture," in *IEEE Personal Indoor and Mobile Radio Communications (PIMRC)*, 2012, pp. 1534–1539.

[40] X. Hou, P. Deshpande, and S. R. Da, "Moving bits from 3g to metro-scale WiFi for vehicular network access: An integrated transport layer solution," in *IEEE International Conference on Network Protocols (ICNP)*, Vancouver, Canada, Oct. 2011, pp. 353–362.

[41] M. A. P. Gonzalez, T. Higashino, and M. Okada, "Radio access considerations for data offloading with multipath tcp in cellular / wifi networks," in *International Conference on Information Networking (ICOIN)*, Bangkok, Jan. 2013, pp. 680–685.

[42] S. Nirjon, A. Nicoara, C.-H. Hsu, J. Singh, and J. Stankovic, "Multi-nets: Policy oriented real-time switching of wireless interfaces on mobile devices," in *IEEE Real Time and Embedded Technology and Applications Symposium*, Beijing, China, Apr. 2012, pp. 251–260.

[43] A. Rahmati, C. Shepard, C. Tossell, L. Zhong, P. Kortum, A. Nicoara, and J. Singh, "Seamless tcp migration on smartphones without network support," *IEEE Transactions on Mobile Computing*, vol. 13, no. 3, pp. 678–692, 2014.

[44] S.-S. Kang and M. W. Mutka, "A mobile peer-to-peer approach for multimedia content sharing using 3g/wlan dual mode channels," *Wireless Communications and Mobile Computing*, vol. 5, no. 6, pp. 633–645, 2005.

[45] D. Zhu and M. Mutka, "Cooperation among peers in an ad hoc network to support an energy efficient IM service," *Pervasive and Mobile Computing*, vol. 4, no. 3, pp. 335 – 359, 2008.

[46] M.-F. Leung and S.-H. Chan, "Broadcast-based peer-to-peer collaborative video streaming among mobiles," *IEEE Transactions on Broadcasting*, vol. 53, no. 1, pp. 350–361, March 2007.

[47] M. Stiemerling and S. Kiesel, "Cooperative p2p video streaming for mobile peers," in *IEEE Computer Communications and Networks (ICCCN)*, Zurich, Switzerland, Aug 2010, pp. 1–7.

[48] S. Ioannidis, A. Chaintreau, and L. Massoulié, "Optimal and scalable distribution of content updates over a mobile social network," in *IEEE INFOCOM*, Rio de Janeiro, Brazil, Apr. 2009.

[49] B. Han, P. Hui, V. S. A. Kumar, M. V. Marathe, J. Shao, and A. Srinivasan, "Mobile data offloading through opportunistic communications and social participation," *IEEE Transactions on Mobile Computing*, vol. 11, no. 5, pp. 821–834, May 2012.

[50] Y. Li, G. Su, P. Hui, D. Jin, L. Su, and L. Zeng, "Multiple mobile data offloading through delay tolerant networks," in *ACM CHANTS*, Las Vegas, NV, Sep. 2011.

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

[52] Y. Chuang and K.-J. Lin, "Cellular traffic offloading through community-based opportunistic dissemination," in *IEEE Wireless Communications and Networking Conference (WCNC)*, Shanghai, China, Apr. 2012, pp. 3188–3193.

[53] P. Baier and K. Rothermel, "TOMP: Opportunistic traffic offloading using movement predictions," in *IEEE Conference on Local Computer Networks (LCN)*, Oct. 2012.

[54] P. Karunakaran, H. Bagheri, and M. Katz, "Energy efficient multicast data delivery using cooperative mobile clouds," in *18th European Wireless ConferenceEuropean Wireless*, April 2012, pp. 1–5.

[55] S. Hua, Y. Guo, Y. Liu, H. Liu, and S. Panwar, "Scalable video multicast in hybrid 3g/ad-hoc networks," *IEEE Transactions on Multimedia*, vol. 13, no. 2, pp. 402–413, April 2011.

[56] H. Seferoglu, L. Keller, B. Cici, A. Le, and A. Markopoulou, "Cooperative video streaming on smartphones," in *Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, Monticello, IL, Sept 2011, pp. 220–227.

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

[58] M. Ramadan, L. E. Zein, and Z. Dawy, "Implementation and evaluation of cooperative video streaming for mobile devices," in *IEEE Personal,*

Indoor and Mobile Radio Communications, (PIMRC), Cannes, France, Sept 2008, pp. 1–5.

[59] S. Sharafeddine, K. Jahed, N. Abbas, E. Yaacoub, and Z. Dawy, "Exploiting multiple wireless interfaces in smartphones for traffic offloading," in *Black Sea Conference on Communications and Networking (BlackSeaCom)*, July 2013, pp. 142–146.

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

[61] L. Al-Kanj, Z. Dawy, and E. Yaacoub, "Energy-aware cooperative content distribution over wireless networks: Design alternatives and implementation aspects," *IEEE Communications Surveys Tutorials*, vol. 15, no. 4, pp. 1736–1760, Fourth 2013.

[62] H. Luo, X. Meng, R. Ramjee, P. Sinha, and L. Li, "The Design and Evaluation of Unified Cellular and Ad-Hoc Networks," *IEEE Transactions on Mobile Computing*, vol. 6, no. 9, pp. 1060–1074, Sep. 2007.

[63] C. Mayer and O. Waldhorst, "Offloading infrastructure using delay tolerant networks and assurance of delivery," in *IFIP Wireless Days (WD)*, Niagara Falls, ON, Oct. 2011, pp. 1–7.

[64] J. Whitbeck, Y. Lopez, J. Leguay, V. Conan, and M. D. de Amorim, "Push-and-track: Saving infrastructure bandwidth through opportunistic forwarding," *Pervasive and Mobile Computing*, vol. 8, no. 5, pp. 682–697, Oct. 2012.

[65] F. Rebecchi, M. D. de Amorim, and V. Conan, "DROiD: Adapting to individual mobility pays off in mobile data offloading," in *IFIP Networking*, Trondheim, Norway, Jun. 2014.

[66] L. Valerio, R. Bruno, and A. Passarella, "Adaptive data offloading in opportunistic networks through an actor-critic learning method," in *ACM CHANTS*, Maui, HI, Sep. 2014.

[67] H. Izumikawa and J. Katto, "RoCNet: Spatial mobile data offload with user-behavior prediction through delay tolerant networks," in *IEEE Wireless Communications and Networking Conference (WCNC)*, Shanghai, China, 2013, pp. 2196–2201.

[68] M. Pitkanen, T. Karkkainen, and J. Ott, "Opportunistic web access via wlan hotspots," in *IEEE International Conference on Pervasive Computing and Communications (PerCom)*, Mannheim, Germany, Apr. 2010, pp. 20–30.

[69] A. Petz, A. Lindgren, P. Hui, and C. Julien, "MADServer: A server architecture for mobile advanced delivery," in *ACM CHANTS*, Istanbul, Turkey, 2012, pp. 17–22.

[70] K. Doppler, M. Rinne, C. Wijting, C. B. Ribeiro, and K. Hugl, "Device-to-device communication as an underlay to lte-advanced networks," *IEEE Communications Magazine*, vol. 47, no. 12, pp. 42–49, 2009.

[71] B. Raghothaman, E. Deng, R. Pragada, G. Sternberg, T. Deng, and K. Vanganuru, "Architecture and protocols for LTE-based device to device communication," in *International Conference on Computing, Networking and Communications (ICNC)*, San Diego, CA, Jan. 2013, pp. 895 – 899.

[72] K. J. Zou, M. Wang, K. W. Yang, J. Zhang, W. Sheng, Q. Chen, and X. You, "Proximity discovery for device-to-device communications over a cellular network," *IEEE Communications Magazine*, vol. 52, no. 6, pp. 98–107, 2014.

[73] M. J. Yang, S. Y. Lim, H. J. Park, and N. H. Park, "Solving the data overload: Device-to-device bearer control architecture for cellular data offloading," *Vehicular Technology Magazine, IEEE*, vol. 8, no. 1, pp. 31–39, Mar. 2013.

[74] M. Zulhasnine, C. Huang, and A. Srinivasan, "Efficient resource allocation for device-to-device communication underlaying lte network," in *IEEE 6th International Conference on Wireless and Mobile Computing, Networking and Communications (WiMob)*, Niagara Falls, ON, Oct. 2010, pp. 368–375.

[75] C.-H. Yu, K. Doppler, C. B. Ribeiro, and O. Tirkkonen, "Resource sharing optimization for device-to-device communication underlaying cellular networks," *IEEE Transactions on Wireless Communications*, vol. 10, no. 8, pp. 2752–2763, 2011.

[76] F. Malandrino, C. Casetti, and C.-F. Chiasserini, "A fix-and-relax model for heterogeneous lte-based networks," in *IEEE 21st International Symposium on Modeling, Analysis & Simulation of Computer and Telecommunication Systems (MASCOTS)*, San Francisco, CA, Aug. 2013, pp. 308–312.

[77] M. Hasan, E. Hossain, and D. Kim, "Resource allocation under channel uncertainties for relay-aided device-to-device communication underlaying lte-a cellular networks," *IEEE Transactions on Wireless Communications*, vol. 13, no. 4, pp. 2322 – 2338, Apr. 2014.

[78] Y. Li, T. Wu, P. Hui, D. Jin, and S. Chen, "Social-aware d2d communications: qualitative insights and quantitative analysis," *IEEE Communications Magazine*, vol. 52, no. 6, pp. 150–158, 2014.

[79] Y. Li, Z. Wang, D. Jin, and S. Chen, "Optimal mobile content downloading in device-to-device communication underlaying cellular networks," *IEEE Transactions on Wireless Communications*, vol. 13, no. 7, pp. 3596 – 3608, Jul. 2014.

[80] E. Yaacoub, L. Al-Kanj, Z. Dawy, S. Sharafeddine, F. Filali, and A. Abu-Dayya, "A utility minimization approach for energy-aware cooperative content distribution with fairness constraints," *Transactions on Emerging Telecommunications Technologies*, vol. 23, no. 4, pp. 378–392, 2012.

[81] Http://www.transferjet.org/, "TransferJet," accessed: 2013-08-21.

[82] Http://wirelessgigabitalliance.org/, "Wireless Gigabyte (WiGig) Alliance," accessed: 2013-08-21.

[83] X. Wu, S. Tavildar, S. Shakkottai, T. Richardson, J. Li, R. Laroia, and A. Jovicic, "Flashlinq: A synchronous distributed scheduler for peer-to-peer ad hoc networks," *IEEE/ACM Transaction on Networking*, vol. 21, no. 4, pp. 1215–1228, Aug. 2013.

[84] 3GPP, "3GPP TSG SA: Feasibility Study for Proximity Services (ProSe) (Rel. 12)," 2012.

[85] S. Wietholter, M. Emmelmann, R. Andersson, and A. Wolisz, "Performance evaluation of selection schemes for offloading traffic to IEEE 802.11 hotspots," in *IEEE International Conference on Communications (ICC)*, Ottawa, Canada, jun 2012, pp. 5423–5428.

[86] 3GPP, "3GPP TS 24.312: Access Network Discovery and Selection Function (ANDSF) Management Object (MO) (Rel. 10)," 2011.

[87] ——, "3GPP TR 23.829: Local IP Access and Selected IP Traffic Offload (LIPA-SIPTO) (Rel. 10)," 2011.

[88] ——, "3GPP TS 23.261: IP flow mobility and seamless Wireless Local Area Network (WLAN) offload (Rel. 10)," 2011.

[89] K. Samdanis, T. Taleb, and S. Schmid, "Traffic Offload Enhancements for eUTRAN," *IEEE Communications Surveys & Tutorials*, vol. 14, no. 3, pp. 884–896, 2012.

[90] C. B. Sankaran, "Data Offloading Techniques in 3GPP Rel-10 Networks: A Tutorial," *IEEE Communications Magazine*, vol. 50, no. 6, pp. 46–53, 2012.

[91] R. Draves and D. Thaler, "Default router preferences and more-specific routes," *RFC 4191*, 2005.

[92] R.Stewart, "Stream control transmission protocol," *RFC 4960*, 2007.

[93] A. Ford, C. Raiciu, M. Handley, and O. Bonaventure, "TCP Extensions for Multipath Operation with Multiple Addresses," *RFC 6824*, 2013.

[94] "MultiPath TCP -Linux Kernel Implementation," accessed: 2014-06-17. [Online]. Available: http://multipath-tcp.org/pmwiki.php/Users/Android

[95] I. van Beijnum, "Multipath TCP lets Siri seamlessly switch between Wi-Fi and 3G/LTE," accessed: 2014-06-17. [Online]. Available: http://arstechnica.com/apple/2013/09/multipath-tcp-lets-siri-seamlessly-switch-between-wi-fi-and-3glte/

[96] M. Dohler, D.-E. Meddour, S. M. Senouci, and A. Saadani, "Cooperation in 4g - hype or ripe?" *IEEE Technology and Society Magazine*, vol. 27, no. 1, pp. 13–17, Spring 2008.

[97] P. Gupta and P. Kumar, "The capacity of wireless networks," *IEEE Transactions on Information Theory*, vol. 46, no. 2, pp. 388–404, Mar 2000.

[98] S. Weber, J. Andrews, and N. Jindal, "An overview of the transmission capacity of wireless networks," *IEEE Transactions on Communications*, vol. 58, no. 12, pp. 3593–3604, December 2010.

[99] A. Ozgur, O. Leveque, and D. Tse, "Hierarchical cooperation achieves optimal capacity scaling in ad hoc networks," *IEEE Transactions on Information Theory*, vol. 53, no. 10, pp. 3549–3572, Oct 2007.

[100] J. Liu, Y. Kawamoto, H. Nishiyama, N. Kato, and N. Kadowaki, "Device-to-device communications achieve efficient load balancing in LTE-advanced networks," *IEEE Wireless Communications*, vol. 21, no. 2, pp. 57–65, April 2014.

[101] T. Doumi, M. Dolan, S. Tatesh, A. Casati, G. Tsirtsis, K. Anchan, and D. Flore, "LTE for public safety networks," *IEEE Communications Magazine*, vol. 51, no. 2, pp. 106–112, 2013.

[102] D. Feng, L. Lu, Y. Yuan-Wu, G. Ye Li, S. Li, and G. Feng, "Device-to-device communications in cellular networks," *IEEE Communications Magazine*, vol. 52, no. 4, pp. 49–55, 2014.

[103] L. Wei, R. Q. Hu, Y. Qian, and G. Wu, "Enable device-to-device communications underlaying cellular networks: challenges and research aspects," *IEEE Communications Magazine*, vol. 52, no. 6, pp. 90–96, 2014.

[104] A. Chaintreau, P. Hui, J. Crowcroft, C. Diot, R. Gass, and J. Scott, "Impact of human mobility on opportunistic forwarding algorithms," *IEEE Transactions on Mobile Computing*, vol. 6, no. 6, pp. 606–620, June 2007.

[105] T. Karagiannis, J.-Y. Le Boudec, and M. Vojnovic, "Power law and exponential decay of intercontact times between mobile devices," *IEEE Transactions on Mobile Computing*, vol. 9, no. 10, pp. 1377–1390, Oct 2010.

[106] V. Conan, J. Leguay, and T. Friedman, "Characterizing pairwise inter-contact patterns in delay tolerant networks," in *ACM Autonomics*, Rome, Italy, 2007, pp. 19:1–19:9.

[107] A. Passarella and M. Conti, "Analysis of individual pair and aggregate intercontact times in heterogeneous opportunistic networks," *IEEE Transactions on Mobile Computing*, vol. 12, no. 12, pp. 2483–2495, Dec 2013.

[108] A. Tatar, T. Phe-Neau, M. D. de Amorim, V. Conan, and S. Fdida, "Beyond contact predictions in mobile opportunistic networks," in *IEEE Wireless On-demand Network Systems and Services (WONS)*, April 2014, pp. 65–72.

[109] Y. Go, Y. G. Moon, G. Nam, and K. S. Park, "A disruption-tolerant transmission protocol for practical mobile data offloading," in *ACM international workshop on Mobile Opportunistic Networks*, Zurich, Switzerland, 2012, pp. 61–68.

[110] E. Hossain, G. Chow, V. C. Leung, R. D. McLeod, J. Mii, V. W. Wong, and O. Yang, "Vehicular telematics over heterogeneous wireless networks: A survey," *Computer Communications*, vol. 33, no. 7, pp. 775 – 793, 2010.

[111] K. Fall and S. Farrell, "DTN: an architectural retrospective," *IEEE Journal on Selected Areas in Communications*, vol. 26, no. 5, pp. 828–836, Jun. 2008.

[112] N. Golrezaei, A. Dimakis, and A. Molisch, "Scaling behavior for device-to-device communications with distributed caching," *IEEE Transactions on Information Theory*, vol. 60, no. 7, pp. 4286 – 4298, Jul. 2014.

[113] M. Cha, H. Kwak, P. Rodriguez, Y.-Y. Ahn, and S. Moon, "I tube, you tube, everybody tubes: analyzing the world's largest user generated content video system," in *ACM SIGCOMM conference on Internet measurement*, San Diego, CA, 2007, pp. 1–14.

[114] D. Lecompte and F. Gabin, "Evolved multimedia broadcast/multicast service (eMBMS) in LTE-advanced: overview and rel-11 enhancements," *IEEE Communication Magazine*, vol. 50, no. 11, pp. 68–74, Nov. 2012.

[115] F. Rebecchi, M. D. de Amorim, and V. Conan, "Flooding data in a cell: Is cellular multicast better than device-to-device communications?" in *ACM CHANTS*, Maui, HI, Sep. 2014.

[116] E. Biondi, C. Boldrini, A. Passarella, and M. Conti, "Optimal duty cycling in mobile opportunistic networks with end-to-end delay guarantees," in *European Wireless*, Barcelona, Spain, May 2014.

[117] E. Biondi, C. Boldrini, M. Conti, and A. Passarella, "Duty cycling in opportunistic networks: the effect on intercontact times," in *The 17th ACM International Conference on Modeling, Analysis and Simulation of Wireless and Mobile Systems (ACM MSWiM 2014)*, Montreal, Canada, Sep. 2014.

[118] L. Pelusi, A. Passarella, and M. Conti, "Opportunistic networking: data forwarding in disconnected mobile ad hoc networks," *IEEE Communications Magazine*, vol. 44, no. 11, pp. 134–141, 2006.

[119] C. Boldrini and A. Passarella, "Data dissemination in opportunistic networks," *Mobile Ad Hoc Networking: Cutting Edge Directions,*, pp. 453–490, 2013.

[120] A. Vahdat and D. Becker, "Epidemic routing for partially connected ad hoc networks," Technical Report CS-200006, Duke University, Tech. Rep., 2000.

[121] T. Spyropoulos, K. Psounis, and C. S. Raghavendra, "Spray and wait: An efficient routing scheme for intermittently connected mobile networks," in *ACM SIGCOMM Workshop on Delay-tolerant Networking*, Philadelphia, PA, 2005, pp. 252–259.

[122] A. Lindgren, A. Doria, and O. Schelén, "Probabilistic routing in intermittently connected networks," *ACM SIGMOBILE mobile computing and communications review*, vol. 7, no. 3, pp. 19–20, 2003.

[123] J. Burgess, B. Gallagher, D. Jensen, and B. N. Levine, "Maxprop: Routing for vehicle-based disruption-tolerant networks." in *IEEE INFOCOM*, vol. 6, 2006, pp. 1–11.

[124] R. Groenevelt, P. Nain, and G. Koole, "The message delay in mobile ad hoc networks," *Performance Evaluation*, vol. 62, no. 14, pp. 210 – 228, 2005.

[125] X. Zhang, G. Neglia, J. Kurose, and D. Towsley, "Performance modeling of epidemic routing," *Computer Networks*, vol. 51, no. 10, pp. 2867 – 2891, 2007.

[126] P. Jacquet, B. Mans, and G. Rodolakis, "Information propagation speed in mobile and delay tolerant networks," *IEEE Transactions on Information Theory*, vol. 56, no. 10, pp. 5001–5015, Oct 2010.

[127] E. Baccelli, P. Jacquet, B. Mans, and G. Rodolakis, "Highway vehicular delay tolerant networks: Information propagation speed properties," *IEEE Transactions on Information Theory,*, vol. 58, no. 3, pp. 1743–1756, 2012.

[128] V. Vukadinović and G. Karlsson, "Spectral efficiency of mobility-assisted podcasting in cellular networks," in *ACM International Workshop on Mobile Opportunistic Networking (MobiOpp)*, Pisa, Italy, 2010, pp. 51–57.

[129] S. Busanelli, F. Rebecchi, M. Picone, N. Iotti, and G. Ferrari, "Cross-network information dissemination in vehicular ad hoc networks (VANETs): Experimental results from a smartphone-based testbed," *MDPI Future Internet*, vol. 5, no. 3, pp. 398–428, 2013.

[130] NS-2, "Network simulator," http://nsnam.isi.edu/nsnam/index.php.

[131] A. Keränen, J. Ott, and T. Kärkkäinen, "The one simulator for DTN protocol evaluation," in *International Conference on Simulation Tools and Techniques (Simutools)*, Rome, Italy, 2009.

[132] NS-3, "Network simulator," http://www.nsnam.org.

[133] T. Camp, J. Boleng, and V. Davies, "A survey of mobility models for ad hoc network research," *Wireless communications and mobile computing*, vol. 2, no. 5, pp. 483–502, 2002.

[134] F. Ekman, A. Keränen, J. Karvo, and J. Ott, "Working day movement model," in *ACM SIGMOBILE workshop on Mobility models*, 2008, pp. 33–40.

[135] D. Kotz and T. Henderson, "Crawdad: A community resource for archiving wireless data at dartmouth," *Pervasive Computing, IEEE*, vol. 4, no. 4, pp. 12–14, 2005.

[136] M. Sammarco, N. Belblidia, Y. Lopez, M. D. de Amorim, L. H. M. Costa, and J. Leguay, "Pepit: Opportunistic dissemination of large contents on android mobile devices," in *ACM MobiOpp*, Zurich, Switzerland, 2012, pp. 79–80.

[137] J. M. Chapin and W. H. Lehr, "Mobile broadband growth, spectrum scarcity, and sustainable competition," in *TPRC*, 2011, pp. 1–36.

[138] "Mobile data offload for 3G networks," White Paper, IntelliNet Technologies, 2011.

[139] M. Yavuz, F. Meshkati, S. Nanda, A. Pokhariyal, N. Johnson, B. Raghothaman, and A. Richardson, "Interference Management and Performance Analysis of UMTS/HSPA+ Femtocells," *IEEE Communications Magazine*, vol. 47, no. September, pp. 102–109, 2009.

[140] D. Calin, H. Claussen, and H. Uzunalioglu, "On femto deployment architectures and macrocell offloading benefits in joint macro-femto deployments," *IEEE Communications Magazine*, vol. 48, no. 1, pp. 26–32, Jan. 2010.

[141] J. Gora and T. E. Kolding, "Deployment aspects of 3g femtocells," in *International Symposium on Personal, Indoor and Mobile Radio Communications*, Tokyo, Japan, Sep. 2009, pp. 1507–1511.

[142] L. Hu, C. Coletti, N. Huan, P. Mogensen, and J. Elling, "How much can wi-fi offload? a large-scale dense-urban indoor deployment study," in *IEEE Vehicular Technology Conference (VTC Spring)*, Yokohama, Japan, May 2012, pp. 1–6.

[143] D. Karvounas, A. Georgakopoulos, D. Panagiotou, V. Stavroulaki, K. Tsagkaris, and P. Demestichas, "Opportunistic exploitation of resources for improving the energy-efficiency of wireless networks," *IEEE International Conference on Communications (ICC)*, pp. 5746–5750, Jun. 2012.

[144] V. Chandrasekhar, J. G. Andrews, and A. Gatherer, "Femtocell networks: A survey," *IEEE Communications Magazine*, vol. 46, no. 9, pp. 59–67, Sep. 2008.

[145] J. G. Andrews, H. Claussen, M. Dohler, S. Rangan, and M. C. Reed, "Femtocells: Past, Present, and Future," *IEEE Journal on Selected Areas in Communications*, vol. 30, no. 3, pp. 497–508, Apr. 2012.

[146] J. Hoadley and P. Maveddat, "Enabling small cell deployment with HetNet," *IEEE Wireless Communications*, vol. 19, no. 2, pp. 4–5, Apr. 2012.

[147] I. F. Akyildiz, W. Lee, M. C. Vuran, and S. Mohanty, "Next generation/dynamic spectrum access/cognitive radio wireless networks: A survey," *Computer Networks*, vol. 50, no. 13, pp. 2127 – 2159, 2006.

[148] K. Berg and M. Katsigiannis, "Optimal cost-based strategies in mobile network offloading," in *International Conference on Cognitive Radio Oriented Wireless Networks*, Stockholm, Sweden, Jun. 2012.

[149] P. Grønsund, O. Grøndalen, and M. Lähteenoja, "Business Case Evaluations for LTE Network Offloading with Cognitive Femtocells," *Elsevier Telecommunications Policy*, vol. 37, no. 2–3, 2013.
[150] H. ElSawy, E. Hossain, and D. I. Kim, "Hetnets with cognitive small cells: user offloading and distributed channel access techniques," *IEEE Communications Magazine*, vol. 51, no. 6, 2013.
[151] A. J. Mashhadi and P. Hui, "Proactive Caching for Hybrid Urban Mobile Networks," University College London, Tech. Rep., 2010.
[152] F. Malandrino, M. Kurant, A. Markopoulou, C. Westphal, and U. Kozat, "Proactive seeding for information cascades in cellular networks," in *IEEE INFOCOM*, Orlando, FL, Mar. 2012, pp. 1719–1727.
[153] M. Fiore, C. Casetti, and C. Chiasserini, "Caching strategies based on information density estimation in wireless ad hoc networks," *IEEE Transactions on Vehicular Technology*, vol. 60, no. 5, pp. 2194–2208, Jun 2011.
[154] X. Zhuo, W. Gao, G. Cao, and Y. Dai, "Win-Coupon: An incentive framework for 3G traffic offloading," in *IEEE International Conference on Network Protocols (ICNP)*, Vancouver, Canada, Oct. 2011.
[155] L. Gao, G. Iosifidis, J. Huang, and L. Tassiulas, "Economics of mobile data offloading," in *IEEE INFOCOM*, Turin, Italy, Apr. 2013, pp. 1–6.

Filippo Rebecchi received B.Sc. and M.Sc. in Telecommunications Engineering from University of Parma, Italy. He is currently a Ph.D. candidate in Computer Science at LIP6 – UPMC Sorbonne Universits, Paris, France, under a CIFRE grant with Thales Communications & Security. He is Co-Chair of the PhD forum at ACM MobiSys 2015. His research activities focus on delay-tolerant networking, mobile data offloading, and mobile 5G networks.

Marcelo Dias de Amorim is a CNRS research director at the computer science laboratory (LIP6) of UPMC Sorbonne Universites, France. His research interests focus on the design and evaluation of mobile networked systems. For more information, visit http://www-npa.lip6.fr/amorim.

Vania Conan is a senior research expert in networking and communications at Thales Communications & Security, in Gennevilliers, France. He received an Engineering Degree (1990) and Ph.D. in Computer Science (1996) from Ecole des Mines de Paris, France, and subsequently an Habilitation Diriger des Recherches degree in Networking from Universit Pierre et Marie Curie, Paris (2012). He is presently head of the networking laboratory at Thales Communications & Security. His current research topics include ad hoc, delay-tolerant and mobile 5G systems and distributed software based network design.

Andrea Passarella (PhD in Comp. Eng. 05) is with IIT-CNR, Italy. He was a Research Associate at the Computer Laboratory, Cambridge, UK. He published 100+ papers on mobile social networks, opportunistic, ad hoc and sensor networks, receiving the best paper award at IFIP Networking 2011 and IEEE WoWMoM 2013, and the Best Short Paper Award at ACM MSWiM 2014. He is Workshops Co-Chair of ACM MobiSys 2015, and was PC Co-Chair of IEEE WoWMoM 2011, Workshops Co-Chair of IEEE PerCom and WoWMom 2010, and Co-Chair of several IEEE and ACM workshops. He is in the Editorial Board of Elsevier Pervasive and Mobile Computing and Inderscience IJAACS. He was Guest Co-Editor of several special sections in ACM and Elsevier Journals. He is the Vice-Chair of the IFIP WG 6.3 Performance of Communication Systems.

Raffaele Bruno is a Researcher at IIT, an Institute of the Italian National Research Council (CNR). He received a Ph.D. in Computer Engineering from the University of Pisa, Italy, in 2003. His current research interests include the design, modelling and performance evaluation of cyber-physical systems, intelligent transportation and smart grids. He has published in journals and conference proceedings more than 70 papers. He served as guest editor for special issues/fast track sections in Elsevier Pervasive and Mobile Computing journal and Elsevier Computer Communications journal. He is currently on the editorial board of Elsevier Computer Communication journal. He was workshop co-chair of IEEE PerSeNS 2006, IEEE MASS-GHS07, IEEE HotMESH 2009-2011, and IEEE SmartVehciles 2014. He was TPC chair of ACM Q2SWinet 2012 and IFIP/IEEE SustainIT 2013.

Marco Conti is a Research Director of the Italian National Research Council. He has published 300+ research papers, and four books, related to design, modelling, and performance evaluation of computer networks, pervasive systems and social networks. He received the Best Paper Award at IFIP TC6 Networking 2011, IEEE ISCC 2012 and IEEE WoWMoM 2013. He is Editor-in-Chief of Computer Communications journal and Associate Editor-in-Chief of Pervasive and Mobile Computing journal. He served as TPC/general chair for several major conferences, including: Networking 2002, IEEE WoWMoM 2005 and 2006, IEEE PerCom 2006 and 2010, ACM MobiHoc 2006 and IEEE MASS 2007.