Exploring Synergy between Communications, Caching, and Computing in 5G-Grade Deployments

Sergey Andreev, Olga Galinina, Alexander Pyattaev, Jiri Hosek, Pavel Masek, Halim Yanikomeroglu, and Yevgeni Koucheryavy

The authors offer a first-hand tutorial on the most recent advances in content-centric networking, emerging user applications, as well as enabling system architectures. They bring into perspective additional important factors, such as user mobility patterns, aggressive application requirements, and associated operator deployment capabilities, to conduct comprehensive system-level analysis.

ABSTRACT

Decisive progress in 5G mobile technology, fueled by a rapid proliferation of computation-hungry and delay-sensitive services, puts economic pressure on the research community to rethink the fundamentals of underlying networking architectures. Along these lines, the first half of this article offers a first-hand tutorial on the most recent advances in content-centric networking, emerging user applications, as well as enabling system architectures. We establish that while significant progress has been made along the individual vectors of communications, caching, and computing, together with some promising steps in proposing hybrid functionalities, the ultimate synergy behind a fully integrated solution is not nearly well understood. Against this background, the second half of this work carefully brings into perspective additional important factors, such as user mobility patterns, aggressive application requirements, and associated operator deployment capabilities, to conduct comprehensive system-level analysis. Furthermore, supported by a full-fledged practical trial on a live cellular network, our systematic findings reveal the most dominant factors in converged 5G-grade communications, caching, and computing layouts, as well as indicate the natural optimization points for system operators to leverage the maximum available benefits.

ADVANCES IN NEXT-GENERATION NETWORKING

With its regulatory timeline set by the International Telecommunication Union (ITU) and known as International Mobile Telecommunications for 2020 (IMT-2020), the standardization of novel fifth generation (5G) communications technology is finally at full speed in 2016. In line with that, the radio access network (RAN) 5G workshop held by the Third Generation Partnership Project (3GPP) in late 2015 defined the radio technology-related research roadmap to meet the proposed IMT-2020 milestones. Consequently, it is now the industry consensus that there will be a new, non-backward-compatible radio access technology as part of the 5G landscape. However, future 5G networks will be much more than yet another radio access standard, but rather an efficient integration of cross-domain networks to offer a sustainable solution for attracting new verticals beyond information and communications technology (ICT) .

Going further, the emerging 5G interface will enable logical *network slices* within a unified communications ecosystem in stark contrast to a legacy collection of dedicated networks for different industries. This end-to-end network slicing should provide improved rate and latency performance, as well as cater for the more efficient use of wireless spectrum. Due to the dynamic and secure network slices, the integrated 5G system can deliver the needed flexibility to many diverse applications and services, thus radically transforming the existing business models. As the industry is currently answering the important questions of how to slice the network appropriately and at what granularity, it is becoming understood that network operators will take advantage of some of the already developed advanced technologies — including software-defined networking (SDN) and network functions virtualization (NFV) — to implement efficient network slicing.

Allowing the dynamic connection and configuration of various components, SDN is a relatively old technology (dating back to the 1990s), but only now do we have the computational power to finally put it to effective use. Building on top of modern high-volume servers, switches, storage, and cloud computing infrastructure, NFV is essentially the cloudification of the network itself, which has the power to virtualize entire classes of network node functions. With the inherent flexibility offered by SDN and NFV, prospective 5G operators may set up services quickly, and move them around as virtual machines in response to dynamic network demands.

Fueled by SDN, NFV, and network slicing, the communications and computing functionalities are beginning to converge within the 5G ecosystem, bringing up the notion of "computing for communications." The latter concept leverages the synergy between the angles of communications and computing by addressing

Sergey Andreev, Olga Galinina, Alexander Pyattaev, and Yevgeni Koucheryavy are with Tampere University of Technology; Jiri Hosek and Pavel Masek are with Brno University of Technology; Halim Yanikomeroglu is with Carleton University

0163-6804/16/$25.00 © 2016 IEEE

the challenge of efficient computation offloading over a wireless channel. With *5G-grade computation offloading*, resource-constrained and energy-hungry user equipment will be able to migrate its heavy computation tasks to (nearby) resourceful servers. Hence, we are witnessing a dramatic paradigm shift from connection-oriented to *content-oriented networking*, which emphasizes data dissemination, storage, and retrieval capabilities, in contrast to past system architectures aimed solely at increased network capacity [1].

However, the aggressive bandwidth requirements of today's and future user applications (which we discuss in the following section) keep pushing cellular network operators to respond promptly with decisive capacity scaling on their deployments. To this end, heterogeneous networks (HetNets) have recently matured as efficient system architectures, where tower-mounted macrocell base stations (BSs) for ubiquitous coverage and network management are complemented across the same geographical area with small cells of different sizes and by various radio access technologies to improve capacity [2]. Hence, contemporary HetNets allow the serving infrastructure to be brought closer to the actual *content prosumers* (producers-consumers) as well as enable network operators to further densify their deployments, especially in urban areas.

While ultra-dense small cell deployments with their improved area spectral efficiency do mitigate the demand in RAN capacity, at the same time they challenge the *backhauling efficiency* [3]. For many operators, deploying high-speed backhaul between increasingly large numbers of small cell BSs and the core network becomes prohibitively expensive. In order to prevent the backhaul capacity from becoming the 5G system bottleneck (especially during peak traffic hours), *caching* at the BSs can be employed for providing content to users instead of straining the backhaul connections.[1] As it is becoming recognized that caching is indeed an efficient solution to alleviate the backhaul capacity requirement — so that relevant data are deployed during off-peak hours in the caches and then accessed during peak traffic hours by the users — there is strong uncertainty around which solutions are most suitable.

Generally, there is a range of alternative architecture choices (some outlined in the course of this article) taking advantage of content reuse to replace backhaul connectivity with storage capabilities. In the end, a system may be desired where small cell BSs with low-rate backhaul but high storage capacity cache the most popular user content. Then backhaul connections can only be utilized to update the cached content at a rate proportional to how the overall demand distribution evolves over time [4]. Further powered by recent progress in affordable memory capacity, transparent caching in strategic locations should allow the speed-up of content distribution as well as improvement of network resource utilization, even when users do not request the same content simultaneously (i.e., leveraging the temporal variability of network traffic). We continue by understanding the origins of such variability.

Emerging User Applications and Services

Transformed Content Acquisition Habits

As powerful smartphones and tablets increasingly permeate the fabric of our lives, humans are also taking more time utilizing them in their daily routines. Today, time spent using smartphones already exceeds web usage on computers. Indeed, a typical image of the last century was to see everyone reading a newspaper while commuting. Presently, this is forgotten, and people have reverted irrevocably to reading news online on their capable handheld devices. With news going mobile and more real time, the underlying business models begin to evolve as well, resulting in shorter publishing cycles and heavier multimedia streaming content. This, in turn, creates *repetitive downloads* of popular content, such as breaking news and online blockbusters, thus leading to excessively redundant data streaming.

Mobile reading as an emerging method of news discovery is but one example of how people share similarity in terms of content semantics and geography. Another example is represented by massive downloads of a new iOS release, which produces the biggest data spikes seen on the Internet so far. Beyond adopting mobile devices for news and collectively acquiring popular files, today's Internet traffic shows a dramatic influx of on-demand video streaming. Global services, such as YouTube and Netflix, are already watched by millions and have in fact spawned a new generation of video *consumption habits*. For instance, Netflix — the world's market leader for subscription video on demand — had over 60 million paying subscribers in the middle of 2015, and this number is predicted to double by the end of 2020.

However, very different from live streaming, with on-demand streaming people do not request the same content simultaneously. This important property, known as *asynchronous content reuse*, means that a few popular files account for the lion's share of the overall traffic. Indeed, there is evidence that typical user demands concentrate on a relatively small library of files [5]. Furthermore, streaming video on demand requests are highly redundant over time and space, which accentuates the need to deeper explore the current statistical traffic properties and changed user content acquisition patterns. Many recent sources reveal that contemporary cellular technology and service providers are not yet capable of delivering seamless, cost-effective, and scalable on-demand video streaming as the underlying Internet architecture is still based on the historic end-to-end model, and we continue by introducing the associated challenges in the following subsection.

Contemporary Technology Challenges Revealed

In current cellular networks, mobile users located at the cell edges already suffer from high energy consumption due to the aggressive transmit powers, and are further disadvantaged by *excessive latency* in acquiring their desired content over a wireless access network. To make matters worse, humans are particularly sensitive to delay and jitter. The large data providers in the market, including Google and Akamai, which own

Mobile reading as an emerging method of news discovery is but one example of how people share similarity in terms of content semantics and geography. Another example is represented by massive downloads of a new iOS release, which produces the biggest data spikes seen on the Internet so far.

[1] Refer to "The Business Case for Caching in 4G LTE Networks," white paper prepared for LSI by Haig Sarkissian

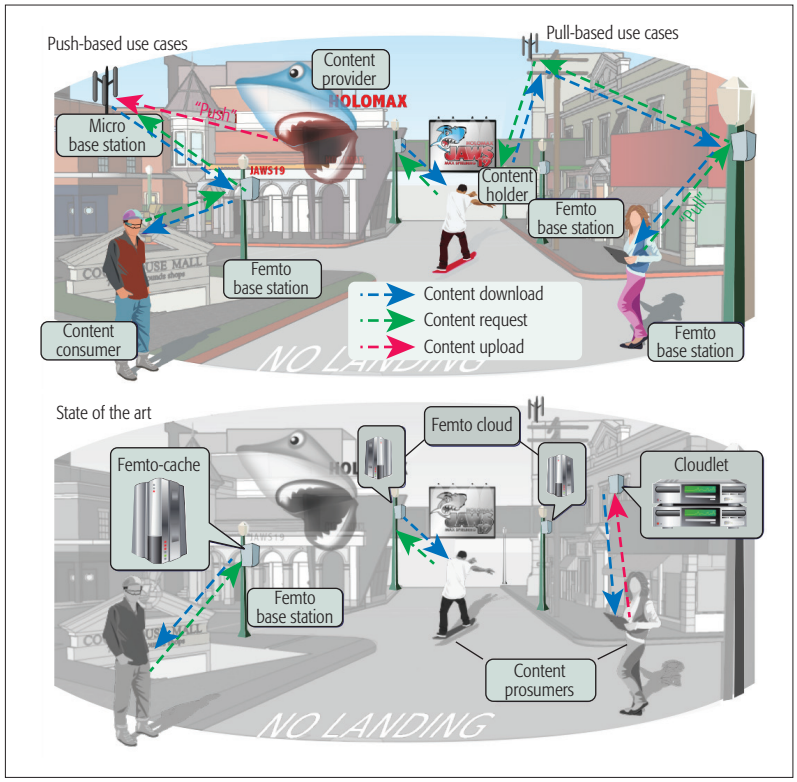

Figure 1. Overview and synergy of emerging 5G-grade solutions.

a multi-billion dollar business built on content delivery networks, recognize that low latency is a key ingredient for the satisfaction of their users. Real-life examples from the Internet confirm that a modest increase in latency can decrease revenues for an Internet service significantly.[2]

To enhance user delay experience, caching popular content at the *wireless edge* (i.e., small cell BSs) could be employed, thus mitigating the disproportion between the available and demanded wireless capacity [6]. However, when deciding exactly what and where to cache, unpredictable *human mobility* may complicate the process profoundly, especially since people increasingly rely on their portable and handheld equipment. Together with understanding mobility, the knowledge of *user location* is equally important to efficiently serve scenarios with co-located devices: passengers with mobile gadgets using public transit services; groups of people in a shopping mall, stadium, and airport; and so on. In the end, it is crucial to consider multiple real-world factors for effective content caching (e.g., popularity distributions, location, velocity, and mobility patterns) in order to accommodate challenging use cases with stringent quality of service (QoS) and computational requirements.

Today, there is already extreme diversity of user applications (with many more to come) that demand extensive computation and continuous processing of the collected data. These include car navigation systems, image processing for electronic games, video processing on smartphones, object recognition on mobile robots, speech synthesis, natural language processing, and wearable computing. However, further development of these novel *5G-grade applications and services* is inherently constrained by the computing efficiency of current user equipment, which is not expected to scale indefinitely due to the fundamental limitations in form factor and battery life. Hence, these novel computation-hungry services will inevitably have to rely on advanced computation offloading capabilities and need to be carefully provisioned in emerging network architectures. We discuss their two characteristic classes in the next subsection.

[2] Refer to "Latency: The Impact of Latency on Application Performance," white paper by Nokia Siemens Networks.

Characteristic Classes of 5G Applications and Services

A particularly challenging use case in the above context, enabled by the latest advances in wearable display and computing technology, is *augmented reality* (AR), which opens the door to truly interactive user experience (Fig. 1). In contrast to virtual reality, AR aims at supplementing the real world, rather than creating an entirely artificial environment. To this end, physical objects in the individual's surroundings become the backdrop and target items for computer-generated annotations, which requires complex real-time calculations. In light of the ongoing content delivery transformation spawning a myriad of computation-heavy and delay-sensitive applications [7], we propose to differentiate between *two large classes* of use cases based on whether the data flow is triggered by a user or by its surrounding network infrastructure.

Pull-Based Use Cases: This category includes user-initiated services (Fig. 1), such as multimedia processing for work and entertainment. Example applications range from editing and creating multimedia content in social networks by amateurs to serving the needs of roaming "deskless" workers up to potentially allowing for hands-free operation (e.g., in medicine, manufacturing, service). Common to all these use cases, the required functionality is "pulled" from its surrounding network by an individual user or a group of users on demand (i.e., similar to the PlanGrid solution for construction engineers).

Push-Based Use Cases: In this category, we collect network-initiated services (Fig. 1), including location-based viral advertising, hazardous environment monitoring, context-aware computing, and mobile AR scenarios. The corresponding applications vary from offering best-effort news and information services to providing real-time capability of object recognition and visualized digital information (spanning the areas of gaming and infotainment, utilities, service and education, guidance, etc.). These use cases commonly assume that the network proactively "pushes" certain services onto a user in a serendipitous fashion.

We proceed further with reviewing the recent progress in enabling network architectures to facilitate these use cases.

Progress in System Architecture Development

Developments in Computing and Networking Infrastructure

In recent years, cloud computing (CC) has become widely recognized as the state-of-the-art computing infrastructure, which has dramatical-

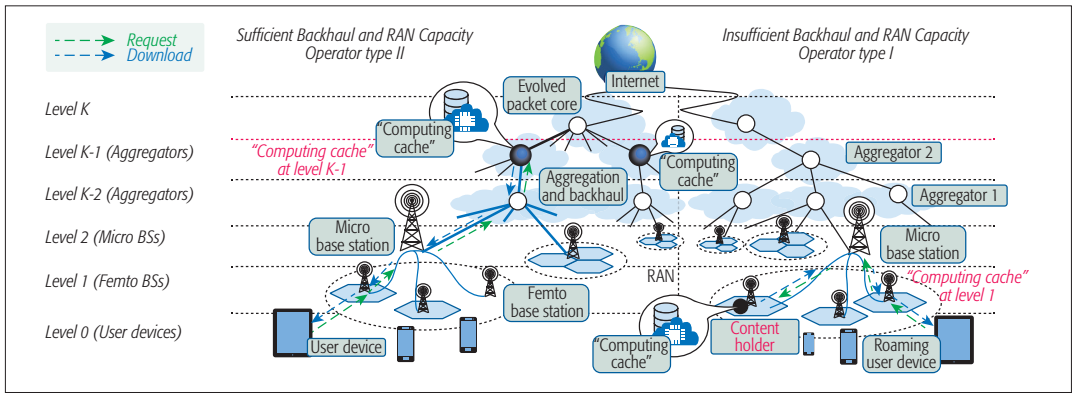

Figure 2. Available system architecture choices and related network "tree" levels.

ly boosted computation offloading capabilities. Building on the virtualization of *computationally intensive processing*, CC leverages the possibility to run multiple operating systems and applications on the same machine(s), while guaranteeing isolation and protection of the programs and their data. This has been instrumental in migrating computation-heavy user applications into the more resourceful cloud and thus has led to enormous economic success. Today, modern cloud service providers enable their users to elastically utilize resources in an on-demand fashion, including infrastructures, platforms, and software [8].

Further integration of CC into the mobile environment has given rise to *mobile CC*, which allows many practical challenges related to performance, environment, and security to be overcome [9]. However, the major limitation of contemporary mobile CC solutions is high *end-to-end latency* experienced during data delivery (including access, transport, and server delay components). To this end, current wireless access networks may introduce extra unwanted latency and often suffer from insufficient throughput due to the regulatory constraints on the available spectrum resources. As a result, cloud service providers are lacking cost-effective means to scale the bandwidth offered to their users, thus making mobile CC services cumbersome to deploy and maintain for handheld and wearable user equipment.

To improve performance on their access networks, mobile operators are taking many decisive steps outlined earlier. However, the ongoing race for a more efficient radio access technology leads to a situation where last-mile wireless connections are sometimes upgraded considerably faster than the corresponding backbone infrastructure. Therefore, some network operators do not have sufficient *backhaul capacity* (deployment type I in Fig. 2), and communicating data to the Internet and back across their deployments takes significant time. A more coherent (but costly) upgrade strategy involves enhancing the backbone together with the last-mile links so as to allow for sustainable growth and support higher traffic loads. This typically requires that the operators deploy more abundant backhaul capacity (deployment type II in Fig. 2). We review the attractive operator choices in more detail further on.

Content Distribution Architectures of Today

An important line of development in networking architecture is related to *content distribution* [10]. Starting from the era of early peer-to-peer (P2P) overlays, contemporary Internet communications pays more attention to the content itself rather than where it is located physically. Powered by Akamai, content delivery networks (CDNs) support anycast methodology by choosing the most appropriate (i.e., topologically close) content replica to achieve self-organized, adaptive, and fault-tolerant content distribution. Furthermore, the concept of information-centric networking (ICN) has developed as a general infrastructure that provides in-network caching so that content is distributed in a scalable, cost-efficient, and secure manner. The essence of these advancements lies in decoupling content from its hosts (or their locations) not at the application layer, but rather at the network layer.

The above is a distinct departure from the conventional client-server architecture, where a client always moves its computational tasks to a more powerful server. However, if the CDN nodes are placed in the core network, there may be insufficient throughput at the wireless edge, and thus a lack of reliable, low-latency service on wireless links. This still remains reality as mobile (cell edge) users often have to run all the processing locally in their devices and then save the execution result on external memory sticks and flash drives for further sharing. Therefore, it has been quickly recognized that the network may need additional architecture options for deploying caches as well as performing data processing.

In light of the above, embedding caching and computing capabilities into heterogeneous wireless networks may achieve significant reduction in response times by mirroring data/service in various locations and in effect bringing the resources (radio access, storage, and computation cycles) closer to where they are actually used [11]. In addition, in-network caching allows *shifting traffic* from peak to off-peak hours, thereby naturally mitigating load variability and reducing congestion. Such distributed local caches typically operate in two phases, content placement (storage) and delivery, but may also require extra system-wide information (e.g., hop count and content popularity distribution), which substantiates the need for *intelligent caching strategies*. As a result, predictive in-network caches support location transparency, facilitate efficient content

CC leverages the possibility to run multiple operating systems and applications on the same machine(s), while guaranteeing isolation and protection of the programs and their data. This has been instrumental in migrating computation-heavy user applications into the more resourceful cloud and thus has led to enormous economic success.

Deploying cloudlets incurs extra costs for their installation and maintenance, and this does not offer any means to handle user mobility. An alternative approach is merging the cloud computing frameworks and the small cell networks as part of another concept named "femto-cloud" computing, where home eNodeBs would support the cloudlet functionality.

distribution, and have the potential to balance the data transmission, storage, and user connectivity costs. The following subsection summarizes the recent progress along these lines.

State of the Art in 5G-Grade "Computing Caches"

Facilitated by the all-IP nature of contemporary 3GPP Long Term Evolution (LTE) cellular networks, two types of locations appear attractive for deploying *5G-grade caches* [12]:

- The Evolved Packet Core (EPC), which consists of the serving gateway (S-GW), the packet data network gateway (P-GW), and the mobility management entity (MME)
- The RAN, which features evolved NodeB (eNodeB) BSs

In some cases, it may also be beneficial to combine the caching functionality with the matching processing power, especially when an application requires repeated bursty access to a remote server or other complex interactions. We name the corresponding architecture node a *computing cache*, which is essentially a virtualized resource available in the 5G network and targeted specifically at remote execution of end-user applications.

Recent literature has been rich in proposing other hybrid deployments of communications, caching, and computing functionalities. Ever since the pioneering work in [13], various options for a mobile device to *cyberforage* by finding surrogate (i.e., helper) servers in the environment have been considered. Proposing to move computation resources closer to the user devices, the concept of a *cloudlet* has emerged offering mobile handsets the possibility to access nearby static resourceful computers, linked to a remote cloud with high-speed wired connections. Within the cloudlet vision, such helper servers would be located in public and commercial spaces (airports, train stations, cafes, etc.) where people congregate casually. Hence, user devices can offload their computations to a nearby server, at low latency and high bandwidth, rather than pushing them to the cloud.

However, deploying cloudlets incurs extra costs for their installation and maintenance, and this does not offer any means to handle user mobility. An alternative approach is merging the CC frameworks and the small cell networks as part of another concept named *femto-cloud* computing, where home eNodeBs would support the cloudlet functionality. Femto-clouds enable a capillary distribution of the CC capabilities, closer to the actual mobile clients. In complement, caching the content library at femtocell stations (so-called *femtocaching* [4]), and even in the mobile devices themselves, has demonstrated particular benefits by alleviating the backhaul requirement in HetNets [14]. Femtocaching has the potential to solve the network scalability challenge by providing user rates with better scaling behavior. In summary, Fig. 1 supports our above discussion on the relationship and synergy between communications, caching, and computing with an overview of the latest research progress.

In addition, a vast body of works has concentrated on development of advanced mobile content caching and delivery techniques, as well as focused on improving network resource utilization. However, all the relevant practical factors need to be taken into account comprehensively to leverage the full synergy of the converged communications, caching, and computing architecture, including the structure of content requests, cost per backhaul connection and operating costs, user mobility control, requirements of running applications, and so on. Inspired by this, in the rest of this article our aim is to offer a unique system-level analysis of such integrated architecture, supported by a live measurement campaign.

Representative Scenarios and Their Evaluation

Evaluation Methodology and Assumptions

Characterizing the converged communications, caching, and computing functionalities, we investigate two representative use cases belonging to the two classes introduced earlier:

- Streaming context-aware AR data (scenario 1, push-based)
- Using a web-based application, such as Adobe Photoshop (PS) cloud (scenario 2, pull-based)

Both example applications require intensive computations, which are cumbersome to run on small-scale user equipment and thus have to be offloaded. To this end, we assume that both storage and computation resources may in principle be located in the LTE network (RAN, EPC, etc.). An end-user session spawns small-size packets: files containing the extracted features for image classification and environment recognition, or PS brush track reports translated into formal commands.

With our detailed system-level simulations, we recreate an urban area of interest (or tracking area) where active users are moving according to a certain random walk model (calibrated with practical measurements in the following subsection). As a reference, we employ fractional Brownian motion with positively-correlated increments (Hurst parameter $H = 0.9$). Hence, our users tend to preserve their movement directions as they keep interacting with the network continuously. In scenario 1, the appropriate *content holder* is determined by the current geographical position of the user, while for scenario 2 a session with a particular content holder has a geometrically distributed duration with the average of 30 min.

In particular, for scenario 1 we consider continual computing and data acquisition (i.e., a user's wearable camera captures the context and annotations are "pushed" by the network) as the user moves across the area of interest. By contrast, for scenario 2, the "pull" requests are sent in ON/OFF fashion, such as when the user is drawing with the PS brush so that remote service is demanded. The period between the requests is taken as 33 ms and 50 ms for scenarios 1 and 2, respectively. For AR, the video frame size that has to be downloaded is 67 kb (i.e., video rate of 2 Mb/s at 30 fps rate), whereas for PS we assume a series of requests during the exponential ON periods with the average of 3 s, and the "silent" exponential OFF periods with the average of 6 s.

Description	Value
Number of levels	5
Femtocell density	8/microcell
Effective user density	560 people/km^2
Femtocell radius	50 m
Microcell radius	200 m
User speed	3 km/h
LTE femtocell capacity	10/10/5 Mb/s
LTE microcell backhaul	100/50/30 Mb/s
Aggregator 1 backhaul	3/1/0.8 Gb/s
Aggregator 2 backhaul	30/8/6 Gb/s
EPC capacity	300/60/38 Gb/s
LTE RAN latency	7 ms
Femto-micro BS latency	3 ms
Small-scale aggregator latency	0.5 ms
Large-scale aggregator latency	2 ms

Table 1. Key system parameters.

From the connectivity perspective, mobile users can communicate with a microcell BS if they are located within its coverage area. Alternatively, users may also connect to one of the femtocell BSs deployed across the tracking area according to a certain stationary repulsion point process. As a characteristic example of femto BS distribution, we consider a Strauss process with the inhibition coefficient 0.9 and the inhibition distance of 90 m. Given that cellular network topology is hierarchical, we further adopt the following abstraction of its structure. We represent the entire operator's network as a forest of trees, where a certain tree (Fig. 2) corresponds to a particular access network "branch," while "leaves" denote the end-user devices. Enumerating the network "tree" levels, we call the user level "level 0," the RAN levels (femto and micro BSs) "level 1" and "level 2," and further on through the aggregation nodes in the backbone network to the EPC (the root level is "level K").

The network structure in our evaluation is instantiated with the typical numbers of descendants expected of a real operator network ({10,10,30,8,var.} starting with EPC descendants and all the way down). The resources of any node are shared fairly between all the active descendants at a lower level (including the user level). In case of a backhaul bottleneck at some level, the maximum possible throughput of every user is decreased proportionally. Furthermore, both data storage and remote processing nodes may in principle be located at any given level of our network topology. Here, we assume that they are always deployed together (co-located) at every node of a certain level, thus mimicking the "computing cache" functionality discussed earlier.

Generally, the end-to-end latency comprises the time to:

- Upload the request τ_{UL}
- Perform the calculations τ_{compute} (either remotely or at the user device)
- Download the final result τ_{DL}

The request timings that form τ_{UL} at all levels are given in Table 1; τ_{DL} and $\tau_{compute}$ depend on the system load and are explained below. To estimate latency, we introduce the rates R_i to traverse the network tree from level $i-1$ to level i (shared between all the active descendants), which are directly related to the system load. Hence, R_1 denotes the individual maximum data rate of a user at the femto BS, R_2 is the rate on a backhaul connection between the femto and the micro BSs, while the rest of the network "tree" edges are wired.

The processing capacity of the computing nodes can be provisioned by the service provider appropriately based on the available funds per user. Therefore, our abstracted model assumes that a certain server at level 1 may process a computational task during 5 ms if there are no other requests. For level i, $i > 1$, since the available computation resource is assumed to scale linearly, the user requests can be served proportionally faster due to additional parallel servers. However, a constant overhead of 10 ms is added to this variable delay regardless.

Supportive Live Trial Implementation

In order to understand real-life user movement behavior and its impact on the performance of our envisioned system, we additionally implemented a full-scale user mobility study. The motivation behind this trial has been to collect live data from end-user devices connected to our open cellular network so as to reveal the effective frequency of serving cell changes by our test population of users. Furthermore, we employed thus collected information as a calibration dataset for a more detailed system-wide evaluation reported at length in this section. The present investigation of practical user mobility builds on our rich hands-on experience acquired during our recent implementation work in [15], which utilizes all of the key functionality expected of 3GPP LTE Release 10.

Our employed LTE testbed (Fig. 3b) is composed of:

- The RAN part, including several small cells
- The EPC part
- The IP multimedia subsystem (IMS) part

To provide a complete and unbiased picture, we also performed supporting experiments in other public mobile LTE networks served by telecommunication operators of the Czech Republic, including Telefonica O2, Vodafone, and T-Mobile. As our test user equipment, we utilized Samsung Galaxy S3 and S4, Jiayu S3 Advanced, as well as Samsung Galaxy Note 4 devices. In addition, we created an assessment tool in Java to collect live information on the cell ID, location of eNodeBs as the location area code (LAC) parameter, received signal strength, and connection latency between a user and the server.

Based on the results of the trial, we are convinced that user mobility is one of the most crucial factors in the present system performance evaluation. Hence, the obtained live measurements were processed to extract the values of the cell residence time (i.e., how long a user spends in one cell before changing it). Then we employed these data in our simulation study discussed in the previous subsection to yield substantiated conclusions on the practical system behavior. We report on our assessment by visualizing the collected results in the form of a daytime scattergram (Fig. 3a) for the cell residence

The motivation behind this trial has been to collect live data from end-user devices connected to our open cellular network so as to reveal the effective frequency of serving cell changes by our test population of users. Furthermore, we employed thus collected information as a calibration dataset for a more detailed system-wide evaluation.

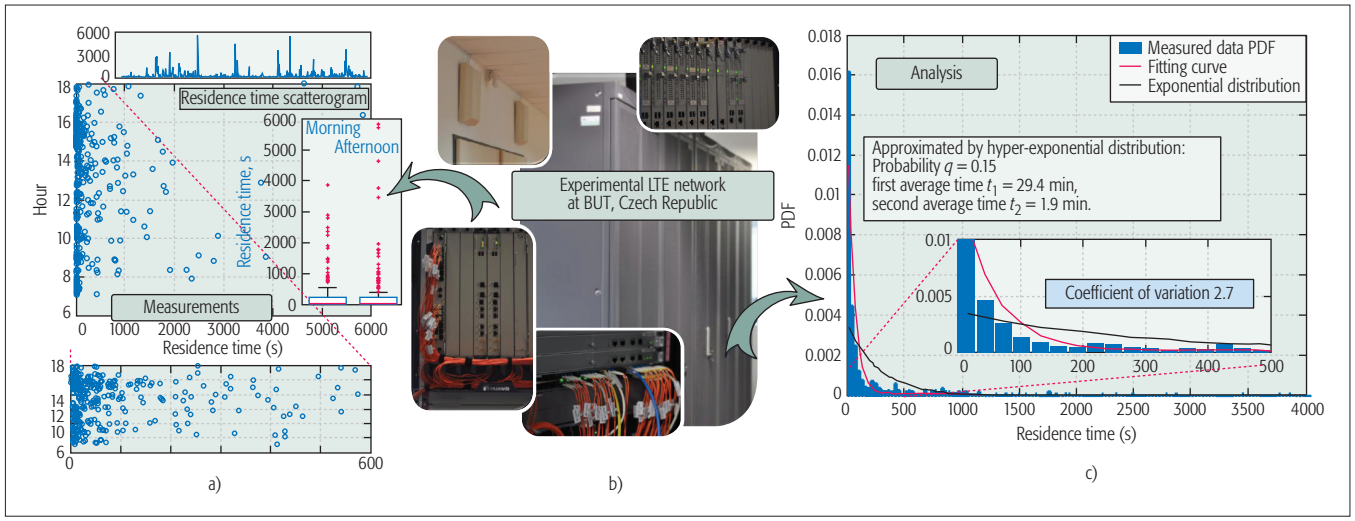

Figure 3. Trial implementation and interpretation of measurements.

times during business hours. Some of the test users demonstrate interesting variations in their mobility patterns, but on the whole the assembled data follow the trend of the unified sample on which we focus in the rest of this discussion.

Another curious finding is that the behavior of the random residence time is not stationary and might alter throughout the day (see box-and-whisker diagram in Fig. 3a for morning and afternoon hours). This is due to the characteristic habits of our test group: the participants tend to move more actively in the morning hours. Further, we note that the empirical probability density function as illustrated in Fig. 3c is very different from the standard exponential distribution, since the coefficient of variation for our sample is much higher than 1. This leads to the need of using more complex fitting options coming from the class of phase-type (PH) distributions, that is, mixtures of distributions. After calibrating with the experimental data, we continue by reporting our most important simulation-based findings.

Selected Numerical Results

To quantify the scaling laws behind the discussed use cases, we consider the system, where the relevant storage and processing functionalities are assumed to be available for a user at a particular "computing cache" node in the network (named "content holder") for both Scenarios 1 and 2, as well as the three different network operator profiles (Fig. 4):

- Sufficient RAN and backhaul capacity (over-provisioned network, where the capacity of a higher-level node equals the total capacity of its subordinate nodes)
- Moderate RAN capacity and insufficient backhaul capacity (capacity of a higher-level node is decreased with respect to the total capacity of its subordinate nodes)
- Insufficient RAN and backhaul capacity (capacity scales down even more severely)

The corresponding deployment parameters are summarized in Table 1.

In real-world networks, computation delay decreases as the associated processing node is placed higher in the network "tree" (due to aggregating multiple computational tasks and allocating more resources). Hence, we expect that offloading to the higher "tree" levels may be more beneficial for the user in that respect. However, the data communications delay is always a non-decreasing function of the "tree" level index, and strongly depends on the current network load as well. Within the two considered scenarios, as user interactions with the network are rather intense, smaller network capacities may have difficulty in supporting the offered traffic load, thus creating an incentive to move the resources closer to the edge. These two conflicting objectives lead to nontrivial results, which also depend on other important factors.

In particular, the computing delay — which is higher at RAN nodes — impacts the total latency in an *underloaded* (well provisioned) network (Figs. 4a, 4d), thus moving the delay-optimal "computing cache" placement point toward the EPC. However, with degraded network capacity (Figs. 4b, 4e and then Figs. 4c, 4f), the computing delay loses its importance to the communications delay, which becomes the dominant factor in determining the user QoS. We confirm that the optimal "computing cache" level in a highly loaded network is at the edge (level 1 or 2, depending on the available computation resources), while in a more lightly loaded system the optimality point shifts to "higher" aggregation nodes. Furthermore, the more intensive data communications is, the sooner the optimal point slides toward the edge.

For our practical setup, we conclude that a typical network hardly copes with the latency requirements on the order of tens of milliseconds. Similarly, a legacy network with insufficient backhaul capacity cannot support the real-time restrictions of our Scenario 1 (AR). However, Scenario 2 (PS) may operate satisfactorily in all the considered deployments, since it is not as delay-critical and throughput-hungry. As seen in Fig. 5d, for the more user-mobility-sensitive Scenario 2, the transport delay at the edge of the network increases due to the fact that the user has to communicate with its original content holder farther away across the network. This certainly has a negative impact on the network

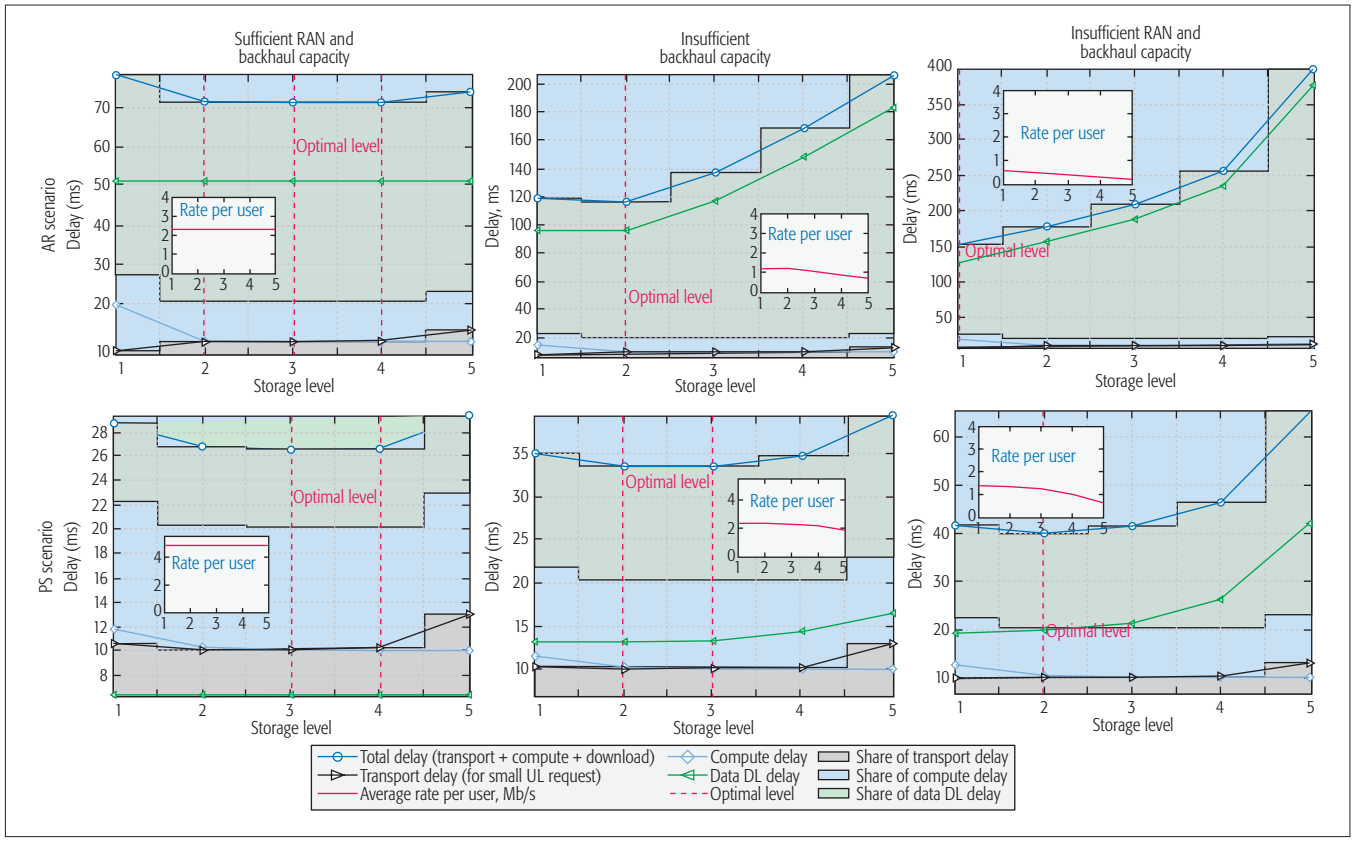

Figure 4. Service delay assessment for the three network types and two scenarios of interest.

load as further demonstrated in Fig. 5b for the moderate-capacity operator deployment. The latter figure highlights the relative load on certain network levels (i.e., the "loaded level") when the "computing cache" is deployed at the "storage level."

In Fig. 5b, Scenario 2 creates more than twice as much extra load at "higher" levels when the "computing cache" is placed at the femtocell level. In contrast, Fig. 5a corresponds to Scenario 1 with its shorter intervals between the changes of the content holder (related to the cell residence times). This results in a stronger correlation of the user's location with that of the helping network node. Note that an additional small portion of load (highlighted separately in the top left corner) is produced by serving users outside of the femtocell coverage. Importantly, due to the higher load of AR and insufficient backhaul capacity, the system bottlenecks (i.e., 99 percent loading) impact the service rate and the network load compared to when the computing cache resides at the edge, which leads to increasing delays (as we have seen in Fig. 4). For a resourceful and well provisioned operator infrastructure, one should expect a "flat" surface of backhaul load.

In summary, we learn that in case of sufficient operator network capacity, the deployment of computing cache nodes at "higher" levels would provide better end-user performance, likely at reduced equipment costs (computing cache nodes may be hosted on already existing server hardware), but then causing a significant load on the RAN (i.e., distribution network). For networks with insufficient backhaul capacity, attractive performance gains are only seen when the computing cache is brought closer to the user, as this mitigates congestion in the distribution network. However, there may be additional deployment costs, which have to be considered when the system is provisioned.

Main Outcomes and Conclusions

As our results conclusively indicate, appropriate deployment of the computing cache functionality in next-generation cellular networks does not have a single universal answer. To adequately quantify the cornerstone questions of what and where to cache, as well as how many computation nodes should be made available and at what level, we considered the realistic provisioning of the emerging 5G-grade applications and services, such as AR and offloaded computation. Not limited to a simple illustration of the attainable gains, the considered scenarios represent the two distinct classes of push-based and pull-based use cases introduced early on in this article.

Our subsequent numerical findings suggest that computing cache nodes have to be deployed by 5G system operators in a manner consistent with a broad range of practical aspects, many of which have been considered in synergy by this work for the first time. The most important of such factors are outlined below.

Network Capabilities: This includes the serving operator infrastructure, from the core down to femtocells, the capacity of all the transit nodes and RAN, the effective coverage ranges and cell density, as well as the computation and caching capabilities together with the cost of their deployment.

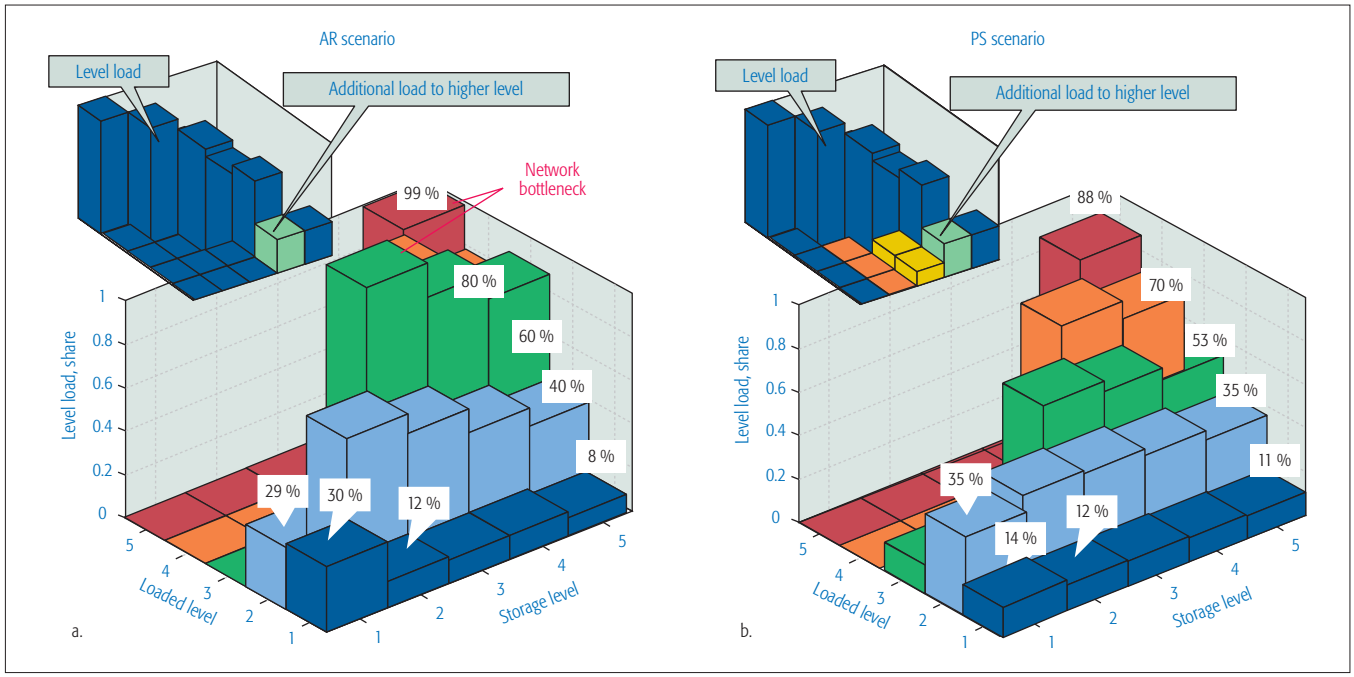

Figure 5. System load analysis (for moderate network capacity).

Application Requirements: Further attention should be paid to the actual service needs, including rate and delay, as well as the structure of computation demand in active periods.

User Dynamics: Other important factors are related to the number of active prosumers, their mobility (including speed and preferred movement patterns), as well as characteristic roaming behavior and cell residence times.

In summary, the analysis conducted in this article reveals that the optimal position of the computing cache nodes is determined by the *application* to be supported, the user *mobility* pattern, as well as the backhaul and RAN *dimensioning*.

More specifically, for the low-bandwidth, high-persistence use cases oriented toward computing (running longer than a user's residence time in a particular area of interest), such as those illustrated by our Photoshop scenario, 5G operators should avoid the use of femtocells as computing cache nodes, contrary to popular belief. This is particularly true in a high-capacity network, where "vertical" backhaul transport delays are minimal, and the impact of handovers as well as low computational power of the femtocells become dominant. The *general guideline* for scenarios of this type is to deploy the computing cache functionality higher up in the distribution network, such that the more powerful computing nodes could be used, and the handover overheads would remain minimal.

However, for the emerging high-bandwidth location-bound services oriented at storage, such as our AR scenario, a different *deployment strategy* is preferred. Due to the properties of the AR use case, it becomes significantly more efficient to place the computing cache as close to the end user as possible. Surprisingly, even in this scenario femtocell-level caching is not always the best option, since considerable handover overheads may still exist in reality. In general, however, for download-oriented cases there is a reasonable motivation for the deployment of cell-level computing caches.

Overall, based on our results, it could be recommended that a 5G operator deploy *different* kinds of computing and caching solutions for various scenarios. In particular, cell-level computing caches should be deployed for bandwidth-hungry applications, whereas the higher-level computing cache positions should be considered for high-persistence, computation-oriented services. Furthermore, if the network is underprovisioned (i.e., its backhaul capacity is reduced considerably), the computing cache deployment choice has little effect on the overall service quality delivered to customers. It is thus imperative that the backhaul capacity between the cells housing the computing caches is sufficient to support handovers. If this requirement is not met, the backhaul quickly becomes overloaded with handover-related traffic, and the QoS levels drop significantly.

Acknowledgments

This work was supported in part by the Academy of Finland and in part by the National Sustainability Program under grant LO1401. For the research, infrastructure of the SIX Center was used. The work of S. Andreev was supported in part by a Postdoctoral Researcher grant from the Academy of Finland and in part by a Jorma Ollila grant from Nokia Foundation. The work of O. Galinina was supported with a personal research grant by the Finnish Cultural Foundation.

References

[1] M. Zhang, H. Luo, and H. Zhang, "A Survey of Caching Mechanisms in Information-Centric Networking," *IEEE Commun. Surveys & Tutorials*, vol. 17, 2015, pp. 1473–99.
[2] S. Andreev *et al.*, "Intelligent Access Network Selection in Converged Multi-Radio Heterogeneous Networks," *IEEE Wireless Commun.*, vol. 21, 2014, pp. 86–96.
[3] Y. Zhou and W. Yu, "Optimized Backhaul Compression for Uplink Cloud Radio Access Network," *IEEE JSAC*, vol. 32, 2014, pp. 1295–1307.

[4] N. Golrezaei *et al.*, "Femtocaching and Device-to-Device Collaboration: A New Architecture for Wireless Video Distribution," *IEEE Commun. Mag.*, vol. 51, 2013, pp. 142–49.
[5] S.-W. Jeon *et al.*, "Wireless Multihop Device-to-Device Caching Networks," *IEEE Trans. Info. Theory*, 2016.
[6] E. Bastug, M. Bennis, and M. Debbah, "Living on the Edge: The Role of Proactive Caching in 5G Wireless Networks," *IEEE Commun. Mag.*, vol. 52, 2014, pp. 82–89.
[7] M. Mirahsan, R. Schoenen, and H. Yanikomeroglu, "HetHetNets: Heterogeneous Traffic Distribution in Heterogeneous Wireless Cellular Networks," *IEEE JSAC*, vol. 33, 2015, pp. 2252–65.
[8] S. Barbarossa, S. Sardellitti, and P. Di Lorenzo, "Communicating while Computing: Distributed Mobile Cloud Computing over 5G Heterogeneous Networks," *IEEE Signal Processing Mag.*, vol. 31, 2014, pp. 45–55.
[9] H. T. Dinh *et al.*, "A Survey of Mobile Cloud Computing: Architecture, Applications, and Approaches," *Wireless Commun. and Mobile Computing*, vol. 13, 2013, pp. 1587–611.
[10] Z. Su and Q. Xu, "Content Distribution over Content Centric Mobile Social Networks in 5G," *IEEE Commun. Mag.*, vol. 53, 2015, pp. 66–72.
[11] J. Hachem, N. Karamchandani, and S. Diggavi, "Content Caching and Delivery over Heterogeneous Wireless Networks," *Proc. IEEE INFOCOM*, 2015.
[12] X. Wang *et al.*, "Cache in the Air: Exploiting Content Caching and Delivery Techniques for 5G Systems," *IEEE Commun. Mag.*, vol. 52, 2014, pp. 131–39.
[13] R. K. Balan *et al.*, "The Case for Cyber Foraging," *Proc. ACM SIGOPS Euro. Wksp.*, 2002, pp. 87–92.
[14] B. Zhou, Y. Cui, and M. Tao, "Stochastic Content-Centric Multicast Scheduling for Cache-Enabled Heterogeneous Cellular Networks," P*roc. ACM CoNEXT Wksp.*, 2015.
[15] A. Pyattaev *et al.*, "3GPP LTE-Assisted Wi-Fi-Direct: Trial Implementation of Live D2D Technology," *ETRI J.*, vol. 37, 2015, pp. 877–87.

Biographies

Sergey Andreev (sergey.andreev@tut.fi) is a senior research scientist in the Department of Electronics and Communications Engineering at Tampere University of Technology (TUT), Finland. He received his Specialist degree (2006) and Cand.Sc. degree (2009), both from St. Petersburg State University of Aerospace Instrumentation, Russia, as well as his Ph.D. degree (2012) from Tampere University of Technology. He has (co-)authored more than 90 published research works on wireless communications, energy efficiency, heterogeneous networking, cooperative communications, and machine-to-machine applications.

Olga Galinina (olga.galinina@tut.fi) is a research scientist in the Department of Electronics and Communications Engineering at TUT. She received her B.Sc. and M.Sc. degrees in applied mathematics from the Department of Applied Mathematics, Faculty of Mechanics and Physics, St. Petersburg State Polytechnical University, Russia as well as the Ph.D. degree from TUT. Her research interests include applied mathematics and statistics, queueing theory and its applications; wireless networking and energy efficient systems, machine-to-machine and device-to-device communication.

Alexander Pyattaev (alexander.pyattaev@tut.fi) is a Ph.D. Candidate in the Department of Electronics and Communications Engineering at TUT. He received his B.Sc. degree from St. Petersburg State University of Telecommunications, Russia, and his M.Sc. degree from TUT. He has publications on a variety of networking-related topics in internationally recognized venues, as well as several technology patents. His primary research interest lies in the area of future wireless networks: shared spectrum access, smart RAT selection, and flexible, adaptive topologies.

Jiri Hosek (hosek@feec.vutbr.cz) received his M.S. and Ph.D. degrees in electrical engineering from the Faculty of Electrical Engineering and Communication at Brno University of Technology (BUT), Czech Republic, in 2007 and 2011, respectively. He is currently a senior researcher at the Department of Telecommunications, BUT. His research work has been concentrated on the design of new communication mechanisms and services for mobile networks. Recently, his research scope also includes the measurement and prediction of end-user satisfaction with mobile data services (QoE).

Pavel Masek (masekpavel@feec.vutbr.cz) received his B.S. and M.S. degrees from the Department of Telecommunication, BUT, in 2011 and 2014, respectively. He is currently pursuing his Ph.D. degree in teleinformatics at the same university. He has publications on a variety of networking-related topics in internationally recognized venues, as well as several technology products. His primary research interest lies in the area of wireless networks: M2M/H2H communication, cellular networks, heterogeneous networking, and data offloading techniques.

Halim Yanikomeroglu (halim@sce.carleton.ca) is a full professor in the Department of Systems and Computer Engineering at Carleton University, Ottawa, Canada. His research interests cover many aspects of wireless technologies with special emphasis on cellular networks. He has coauthored more than 80 IEEE journal papers on wireless technologies. His collaborative research with industry has resulted in about 25 patents (granted and applied). He is a Distinguished Lecturer for the IEEE Communications Society and a Distinguished Speaker for the IEEE Vehicular Technology Society.

Yevgeni Koucheryavy (yk@cs.tut.fi) is a professor and lab director at the Department of Electronics and Communications Engineering of TUT. He received his Ph.D. degree (2004) from TUT. He is the author of numerous publications in the field of advanced wired and wireless networking and communications. His current research interests include various aspects in heterogeneous wireless communication networks and systems, the Internet of Things and its standardization, as well as nanocommunications. He is an Associate Technical Editor of *IEEE Communications Magazine* and an Editor of *IEEE Communications Surveys & Tutorials*.

It is thus imperative that the backhaul capacity between the cells housing the computing caches is sufficient to support handovers. If this requirement is not met, the backhaul quickly becomes overloaded with handover-related traffic, and the QoS levels drop significantly.