

The Building Blocks of the 5th Generation (5G) Wireless Networks for Affordable and Ubiquitous Broadband Wireless Connectivity

Halim Yanikomeroglu

Department of Systems and Computer Engineering
Carleton University
Ottawa, Canada

Keywords: Broadband wireless, cellular networks, 4G/5G, radio access networks, radio resource management

Abstract: Commercial cellular wireless communications have a history of about 25 years. Currently, the wireless industry is busy with the standardization of the 4th generation (4G) cellular networks, namely, LTE/LTE-Advanced and WIMAX 802.16m. The 4G standards are expected to be finalized in the next year or two; the early commercial deployments could start in subsequent years. With 4G networks, wireless internet connectivity will be faster and more affordable which will result in substantial increase in wireless internet usage.

Since the 4G concepts have already moved to the standardization phase, we must begin to work on the building blocks of the following generation (which we refer to as 5G) wireless networks. These networks will facilitate the provision of ubiquitous and affordable broadband (very high speed) wireless connectivity. This paper aims at highlighting some of the concepts and technologies which will facilitate the affordable provision of very high data rates with virtually ubiquitous coverage in 5G wireless networks. We refer to this goal as enabling the 4A paradigm: “any rate, anytime, anywhere, affordable”. In particular, this paper focuses on the coherent integration of advanced radio resource management (RRM) techniques with certain advanced physical layer (PHY) operations in the presence of advanced radio access network (RAN) architectures; we refer to this design principle as the “integrated cross-layer cross-network design”.

I. Motivation

The road towards 4G (4th generation) wireless networks has been the driving force behind a number of global research initiatives in the last few years. However, at the preparation of this paper (December 2009) we have formal documents outlining 4G objectives (such as IMT-Advanced) and tangible standardization efforts towards achieving those objectives (namely, LTE/LTE-Advanced and 802.16j/m). Therefore, 4G is moving from research phase to development phase.

The high-level goal of this paper is to explore the technologies which will facilitate the affordable provision of very high data rates with virtually ubiquitous coverage in beyond-4G (which we refer to as 5G) wireless networks. We refer to this goal as enabling the 4A paradigm: “**any rate, anytime, anywhere, affordable**”.

In the technical level, the paper focuses on the coherent integration of **advanced radio resource management** (RRM) techniques with certain **advanced physical layer** (PHY) operations in the presence of **advanced radio access network** (RAN) architectures; we refer to this design principle as the **integrated cross-layer cross-network design**.

In the coming years we will witness the emergence of extremely high rate applications some of which may require high reliability and/or may be delay-intolerant (such as, telepresence or high-definition video conferencing). Delivery of such applications will constitute a major stress in the network. From the rate perspective alone, one or a handful of users may use the entire capacity offered by a BS (base station).

The cellular industry has so far been fortunate that all the applications it has been providing have been rather easy to handle; this fact contributed substantially to the great commercial success of the cellular industry. The level of difficulty in providing a digital application depends on how much that application is demanding with respect to three factors: rate, reliability, and latency. Digital voice has been the predominant application since the advent of 2G cellular networks. Voice is a relatively low rate (around 10 Kbits/sec) application, which also does not require high reliability. The fact that voice is delay-intolerant does not constitute a major problem, as

our listening is highly tolerant to errors, and thus there does not exist a need for an ARQ (automatic repeat request) mechanism which will not be suitable for a delay-intolerant application.

The more recent SMS (short message sending) application is even an easier type. SMS is very low rate and highly delay-tolerant. SMS requires more reliability than voice, but this is not a problem due to the delay-tolerant nature of SMS (as such, the transmission of an SMS message may be delayed until the channel is good enough, or ARQ can be employed if necessary).

Most recently e-mail has become a common application in cellular networks. Except for downloading large attachments, e-mail is also a relatively low rate application which is delay-tolerant as well. E-mail requires higher reliability; however, like SMS, its delay-tolerance enables the proper operation of mechanisms for achieving high reliability. For those reasons, SMS has been generating substantial revenue for the operators.

A related challenge is the potentially high level of traffic variation (non-uniformity) across the area of network's coverage, which did not previously exist in the context of voice and SMS traffic, but becomes more pronounced with data traffic. If the traffic distribution is known a priori, the variation does not constitute a problem, because a region with a higher traffic density (such as a metropolitan area or a business districts) can be served with a higher density of BSs. However, if this variation becomes highly unpredictable, then creating static BS infrastructure constitutes a challenge. This non-uniformity has not been a serious issue until now; because in voice communications (which is a constant bit rate application) the extent of unpredictable variation has logical limits; for instance, the user density in a hot spot may be at most a few times higher than that in the surrounding areas. But when the applications include the very high rate types, a user with such an application may, temporarily, end up consuming network resources equivalent to hundreds of voice users.

To make the situation even more complicated, future networks will have high traffic fluctuations in the time domain as well. The traffic variation with respect to time has two scales, long and short. The long term variation is related to the time of the day (more traffic in urban areas during the day time and more traffic in suburban areas in evenings); this type of variation exists for all applications including voice. Short-term traffic variations, on the other hand, are observed when many bursty high rate applications coexist at the same time. For instance, a user may initiate a high-definition video-streaming, but then may simply cancel the request as her/his attention moves to some other item during web browsing. The combined actions of many users create very difficult to address randomness in traffic fluctuation. In fact, the way this randomness is addressed in the current wireline networks is by gross over provisioning of resources. This is not an option in wireless networks with finite radio resources.

In summary, the fluctuations in traffic demand (both in space and time) will be at unprecedented levels in future networks, which will result in tremendous stress to wireless operators.

We will put aside, for the time being, the varying levels of reliability and delay sensitivity associated with different traffic types, and focus only on the rate aspect. Now, let us consider a 3-dimensional plot representing the traffic demand in a region where the x- and y-axes show the space (geography) and the z-axis the rate demand at a particular time instant. Next, let us also consider the time evolution of this 3D plot. (As mentioned above, the dynamic range of this plot is expected to be very high, as such, a logarithmic z-axis will be more appropriate). The fundamental question we are facing is, **how can this highly varying traffic demand in the given region be served in the most efficient way?**

II. The 4A Paradigm

The solution of the above posed question will enable, what we call, the 4As paradigm: “**Any rate, anytime, anywhere, affordable**”. “Anytime” and “anywhere” refer to the ubiquity of the coverage in space and time; these have been the buzzwords since 2G networks in the context of voice. As the wireless technologies evolved, networks got sophisticated, and the deployment became denser, the “anytime, anywhere” goal has almost been achieved. “Any rate” refers to the availability of very high rates; this was previously not an issue for voice nor for SMS and e-mail. Clearly, “any rate” is crucial for very high speed internet and other envisioned applications. Providing “any rate” in certain isolated localities will not be difficult (similar to WLAN coverage); however, combining “any rate” with “anytime, anywhere” (which results in ubiquitous very high data rate coverage) creates a great challenge for future wireless networks. It is obvious that the “any rate, anytime, anywhere” combination will be incomplete (and, thus, not meaningful) unless it is “affordable” by the end user. The importance of affordability cannot be stressed enough. Affordability is currently not a major concern for voice

(the flat rate voice plans are quite affordable); for instance, the cell phone penetration is around 80% in Canada, and the penetration is quite high even in developing countries. In contrast, although all cellular operators in Canada provide (moderate rate) wireless internet access, few users actually use such services (perhaps with the exception of basic web browsing) due to the high costs associated (the situation is not very different in other parts of the world as well). It is not difficult to imagine the essentialness of affordability with applications which require rates orders of magnitude higher than today's fastest applications.

The fulfillment of the “any rate, anytime, anywhere, affordable” paradigm is arguably the most important challenge towards the 5G cellular networks. The next obvious question is to identify the key enabling technical concepts for 5G. But first we will state an observation.

A standard such as 4G LTE is developed over a number of years by researchers and engineers (overwhelming majority of whom are PhDs) from very many companies; some of these researchers are among the best experts in industry, with a lot of experience. These experts collectively spend literally tens of thousands of hours during the development of the standard. Moreover, they rely on the research done in their companies as well as the research reported in the open literature which is conducted in universities and research institutes around the world. Likewise, the next-generation cellular standards will be the result of a tremendous amount of collective research effort worldwide. Incremental advances in particular technologies are certainly very important towards enabling the next-generation networks. However no technology, even a breakthrough in a particular area, could single-handedly overcome all the technical challenges. Rather, the combination of a number of advanced technologies will collectively form the backbone of 5G.

With this observation in mind, the aim of the proposed project is not to investigate one single technology in isolation. Instead, the project will first identify some of the most important pieces, and will try to bring those pieces together in the most coherent way. Those crucial pieces will include

- the technologies which have been studied before but have not been implemented on a large scale due to various reasons (such as interference cancellation at the user terminals);
- the state of the art technologies (for instance, those considered for LTE-A and 802.16m); and
- the emerging concepts in the research circles, which look promising but are not mature enough for being considered during the current 4G standardization process.

The main vision in this project is a well-integrated **advanced PHY** and **advanced RRM** in the presence of an **advanced RAN**. The cross layer integration of PHY and RRM is envisioned at a level more substantial than the current cross-layer design approaches. The advances in many other technologies such as software, integrated circuit, display, and battery will also play critical roles; however, those technologies are outside the scope of this paper.

III. Integrated Cross-Layer Cross-Network Design

A high number of technologies will play important roles in 5G networks. However, at the core of all of these technologies lies the RAN. We aim to develop an advanced RAN architecture augmented with such network elements as multihop relays, coordinated multipoint transmission and reception, network MIMO, distributed antennas, and femto-cells. Advanced RRM strategies and algorithms will play an essential role in making these network elements work coherently. The proposed research will encompass the physical, multiple-access, and networking layers of the protocol stack with a cross-layer design approach.

For any wireless application, a certain amount of received energy is required for each bit of information. As is well known, the received bit energy decreases with increases in data rate, as well as with increases in distance between the transmitter and receiver. Voice does not require high rates; hence, it is not difficult to maintain sufficiently high bit energy even when the UT is far from the BS. But for high end wireless internet applications, the received bit energy will be too weak to detect unless the UT is very close to the BS. It has already been argued that deploying a very high number of BSs to make sure that each UT is almost always sufficiently close to a BS is not a feasible or practical solution.

One concept that has been articulated in the wireless research community in the last ten years is to enrich the network with wireless relays, which are relatively inexpensive transceivers deployed by the wireless operators, as helpers to the BSs. When a UT is too far away from the BS, a close-by relay (which could be mounted on a

lamp post) may pick up the signal, clean it up (remove noise and interference as much as possible) and retransmit the signal to the BS. As such, with little transmit energy the high rate signal may travel a longer distance eventually reaching the BS in two radio hops. The concept of relaying (previously often perceived as too complicated) is finally being adopted in 4G LTE-Advanced standard which is expected to be finalized in about two years; the relay based commercial deployments could start in subsequent years (around 2015).

However, the introduction of the two-hop relaying capability to wireless networks will provide only a temporary relief. The extremely high speed applications of the future will necessitate a high number of relays so that a UT's signal may reach the BS in several hops each with only short distances. Other UTs in the network will also be involved in relaying, in addition to the relays deployed by the operators. Moreover, there will be several types of base-stations (BSs) depending on demand; these BSs are often referred to in the research community according to the size of their service areas: macro-BSs, micro-BSs, pico-BSs, and femto-BSs. UTs will likely be capable of communicating with several of these network elements (BSs and relays) concurrently. The proper operation of this complex wireless mesh network requires a vision very different than what wireless operators are accustomed to; developing this vision is part of this research proposal.

In current wireless networks, BSs make the RRM decisions and inform the UTs; in other words, UTs play a passive role in a centralized decision mechanism. Although an advanced wireless network brings great potential towards enabling broadband connectivity, this potential cannot be realized without efficient RRM algorithms. As the wireless networks become more complex in order to deliver very high speeds, the corresponding RRM decisions also become increasingly complex and intricately interrelated. In such complex networks, centralized RRM becomes totally impractical as it is not possible for the BS to make the right RRM decisions for all UTs and to inform them of these decisions, given that the BS and UTs may be several radio hops apart and that the medium is highly dynamic. Therefore, the UTs ought to make the RRM decisions autonomously. Such a distributed architecture may be a great concern to wireless operators, because if the decisions made by UTs happen to be inefficient or wrong or selfish, this situation may bring down the entire network. In any case, the autonomous/distributed RRM is the only feasible mode of operation in the presence of advanced wireless networks which will enable cost-efficient wireless broadband connectivity. Selfish users can be dealt with properly implemented admission control (traffic shaping).

The difficulty is that the existing RRM methods and algorithms developed during the last two to three decades presume a rather simple wireless network model solely composed of BSs where all decisions are made in a centralized manner. These methods and algorithms are often not amenable to modifications to work in the envisioned complex wireless mesh networks where many decisions have to be made in a distributed/autonomous manner with the required reliability and robustness. In order to develop the appropriate RRM methods and algorithms which will constitute the major building blocks of the future wireless networks, novel theories and fresh ideas are needed. Interestingly, as far as we can see, the sought-after theories already do exist to a great extent; however, since these theories have been developed in other disciplines, wireless researchers are often not familiar with them. It should be mentioned that in recent years, there has been some limited success in bringing in two well-established theories from applied mathematics to the communications research, namely, game theory and optimization theory. These are important developments in the right direction, but more is needed.

There is a clear need to bring ideas and inspirations from the theories of **machine learning** and **artificial intelligence** from computer science, **adaptive control** from engineering, and **game theory** and **optimization** from applied mathematics, and to use these with the existing PHY, RRM, and RAN expertise in our research group in the design and operation of future wireless networks which will enable the envisioned broadband connectivity.

Clearly, RAN and RRM are highly interrelated; next we turn our attention to the RRM and PHY interrelation. There has been a historic disconnect between the PHY and RRM research communities mainly due to the evolution of these research branches. The conceptual roots of many RRM strategies can be traced back to wired networks. Wired network protocols have a long and rich history; conventionally a layered architecture (such as the OSI model) has been adapted with limited interaction between layers. When wireless networks emerged, many resource management ideas got inherited from the then well-established wired networks. However, although the concept of sharing limited resources is common in both wired and wireless networks, two important characteristics of the wireless channel makes the RRM in wireless networks a more important and complicated task: a) Due to the broadcast nature of the wireless channel, the links are tangled. b) The channel variations are highly disparate among the links (due to distance, large-scale fading, and slow-scale fading). Therefore, the decision making process (RRM) is much more crucial in wireless networks, as some decisions

are much better than some others. When we move from the *link* setting to the *network* setting, PHY and RRM become more interrelated; therefore, there has been significant interest in cross-layer design during the last decade. However, most cross-layer design schemes proposed in the contemporary literature bring together PHY and RRM mainly through a look-up table. Within that simplistic view, it is enough for the RRM scheduler to know the spectral efficiencies achieved for all possible UT-to-resource-block assignments; and this can be implemented through a look-up table that converts the SNR (signal-to-noise ratio) to spectral efficiency. Therefore, it is not accidental that RRM researchers do not need to know (and, often do not know) much about PHY, and vice versa! Consequently, it will not be an exaggeration to state that the current cross-layer design philosophy is rather superficial.

As explained in the previous section, it is possible to obtain more performance gains by introducing more and more complexity in each of RRM, PHY, and RAN. However, as also argued, it is not a good strategy to try to meet all the network challenges by focusing only in either RRM, or PHY, or RAN, due to the fact that the resulting architecture or the corresponding algorithms become exceedingly complex with saturating additional returns. The better strategy is to bring together advanced RAN, advanced RRM, and advanced PHY in the most meaningful way as explained next; we call this strategy **integrated cross-layer cross-network design**.

IV. Advanced RAN

The transmission rate can be increased by increasing the MIMO gain and/or the bandwidth and/or the spectral efficiency (μ). Although increasing the rate through the MIMO architecture looks very attractive, it is not feasible to deploy a high number of antennas, especially at the UT, due to a number of practical limitations. Increasing the transmit power to achieve a higher μ is not very profitable due to the logarithmic relation between SNR and μ , in addition to many other prohibitive factors associated with high transmit power levels, once again, especially at the UT. Finally, bandwidth is scarce, and the licensed portion of it is very expensive. Besides, increasing the transmission bandwidth results in a linear decrease in SNR; therefore, even if the available bandwidth happens to be very high, the received power will also have to be very high to guarantee a sufficient SNR. This, in turn, means that high path-losses cannot be tolerated; as such, UT and BS cannot be too far apart. It is clear that the data rate, the dynamics of which are governed by the above equation, is not unbounded in practical scenarios. If higher and higher rates are required in a given area, deploying a dense network of BSs with a very dense channel reuse scheme becomes inevitable.

The envisioned advanced RAN has the following elements (refer to Figure 1):

- Central Stations (CSs)
- Base Stations (BSs)
- Distributed Antenna Ports (DAPs)
- Fixed Relay Stations (FRSs)
- Terminal Relay Stations (TRSs)
- User Terminals (UTs)

BS: This refers to the conventional BS which does not need any further elaboration.

CS: A group of neighbouring BSs may be connected to a CS to facilitate the BS coordination/cooperation; as such, the presence or absence of a CS depends on the nature and extent of the BS coordination. That is, there may not be a need for an explicit CS if the information exchange between the BSs is sufficient (this will be the case if the corresponding BS coordination algorithms are distributed), or, if one of the BSs acts as a CS.

DAP: This can be considered as a low-cost micro- or pico-BS with limited functionality (and low transmit power) wired to a full-fledged BS where many decisions related to a set of DAPs are made. A DAP can be considered as a wired relay; “radio-over-fibre” is also often used in the literature to refer to the concept of deploying limited-functionality micro-/pico-BSs connected to a capable BS. How much functionality a DAP has is a design consideration; at the logical extreme, all the signal-specific processing (including detection) may be performed at the BS, in addition to the RRM decisions, provided that there is sufficient merit to justify this scenario (such as the creation of a network MIMO).

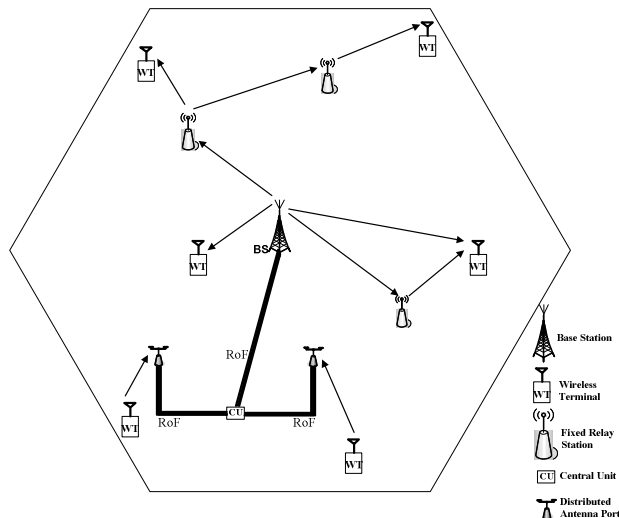


Fig. 1: A typical cell demonstrating interaction among the network entities (DAP, FRS, UT, BS in the envisioned advanced RAN)

FRS: This refers to a relatively low-power relay deployed by the operator, preferably in a strategic location. An FRS, by its very definition, does not need a wired connection to the network (it does need, however, a wired DC source, unless some sort of an alternative energy source (such as solar) is employed). An FRS may communicate with other FRSs and multiple BSs; therefore, the topology is of mesh type.

TRS: This refers to a wireless terminal acting as a relay.

UT: This refers to the terminal of an end user. Some UTs may act as TRSs when need arises.

Comparisons between the RAN elements

- A BS and a DAP have the common feature that they are connected to the backhaul through the wired medium (fibre).
- On the other hand, a BS and a DAP are different in the processing capability and functionality; we perceive DAPs as wired relays with limited functionality (in comparison to full-fledged BSs) in order to achieve cost-effective deployment.
- In contrast, FRSs and TRSs do not have wired connections. An FRS is likely to have much more functionality and capability (signal processing, MIMO, security, RRM, etc.) in comparison to a TRS.
- It is also important to highlight that an FRS has a continuous DC power source from an outlet unlike a TRS which is battery-powered.

The main argument regarding FRSs has been the deployment cost advantage (in comparison to, for instance, micro-BSs) since FRSs do not require the wired backhaul. On the other hand, fiber penetration is relatively high in certain parts of the world (such as South East Asia), and this penetration will be at a much higher level in the years to come. In such cases, it makes sense to utilize low-cost DAPs in addition to FRSs. Smart RRM algorithms will enable the concurrent operation of DAPs and FRSs.

In conventional cellular networks each UT is assigned to one BS; in such networks handoff is an undesirable event implemented through hysteresis to prevent ping-pong effects. Moreover, in the conventional cellular design, a channel is not supposed to be used more than once in a cluster. In the conventional non-CDMA networks, the cluster size is almost always greater than one cell. In order to avoid cell planning, there has been an increasing interest in adopting the single-frequency concept in OFDMA networks as well which will allow the reuse of a channel in every cell (sector) whenever the conditions allow, i.e., the cluster size can be as low as one (as a matter of fact, the resources can be reused in every sector).

We envision a highly ambitious wireless mesh network architecture to support very high data rates with highly bursty and geographically non-uniform traffic patterns. The considered air interface is OFDMA-based. Clearly, the use of the radio and network resources in the most efficient way is of paramount importance. Any fixed assignment or routing association will be inefficient as it will not be able to fully exploit the dynamic conditions in the network. Towards that end, in the envisioned mesh network there is no, or minimal (only

whenever necessary), a priori channel allocations, no a priori radio resource assignments (for instance, opportunistic intra-cell reuse may be possible through power control), and no fixed routing associations; all such decisions are made dynamically and opportunistically (in other words, everything is up for grabs).

A UT's data may be routed through different DAPs or FRSs to a BS or even to different BSs (this mesh architecture may be considered as the evolved version of the CDMA soft handoff concept); moreover, a UT's data over multiple subchannels may be sent through multiple routes to the BS or CS. A great level of diversity gain against shadowing (and, if RAN is advanced enough, against multipath fading as well) can be achieved, and load balancing against congestion (which may occur as a result of the highly variable and bursty traffic) can be attained through opportunistic routing.

It should be noted that the notion of cell becomes rather fuzzy in the articulated advanced RAN. Through the mesh topology, a UT's signal may be routed through FRSs to a further away BS or DAP whenever there is merit (such as avoiding congestion)

Concluding Remarks

We conclude by reflecting, once again, on the fundamental question posed earlier: "How can a very high level of demand, which is also highly varying in time and space, be served in the most efficient way?"

We point out that some of the demand (traffic) will be delay-tolerant; this delay-tolerance can be exploited within the integrated cross-layer cross-network design framework. Even if the demand is more than the supply at a particular instant, this does not create a problem as long as the total demand over a period of T seconds is less than the maximum supply during the same duration. Note that since we introduced time in the discussion, the supply and demand are now represented in bits. Here, without loss of generality, we assume that all the demand can tolerate T seconds of delay (T can be arbitrarily small).

The discussions in this paper suggest the following design principle:

- Obtain the total demand in space and in a time window T seconds.
- Find the absolute minimum number of BSs (N_{\min}) which can handle this demand. Note that although N_{\min} BSs has enough potential supply, in the absence of an advanced RAN, these BSs can satisfy the demand only in very special cases as outline before.
- Deploy N BSs where N is slightly greater than N_{\min} ; then enrich the network with other advanced RAN elements.
- Through the integrated cross-layer cross-network design framework, any spatial and temporal traffic distribution should be handled, *in principle*.

However, the RAN architectures, the RRM protocols, the PHY techniques, and their cross-layer cross-network integration towards facilitating this last bullet are open problems to a great extent.