5G and the Future of IoT
And on its Hardware/Software Impact

Gerhard P. Fettweis
Vodafone Chair, TU Dresden
cfaed, HAEC, 5GLab Germany
Dresden, Germany
gerhard.fettweis@tu-dresden.de

Abstract—We are at the doorstep of 5G wireless communications technology to be defined in standards and then being introduced. Besides the ever increase in data rate the difference over 4G will be defined by additional enabling connectivity. On the one hand, massive Internet of Things (IoT) will become possible. And on the other hand an infrastructure for remote controlling real and virtual objects, the Tactile Internet”, will be enabled. This requires a new approach in architecting hardware/software architectures for digital processing of radio signals as well as applications. Scalability at a new level needs to be created to build a design infrastructure for developing the new application solutions enabled by 5G.

Keywords—5G; IoT; Tactile Internet; MPSoC; SDR; Edge Cloud

I. INTRODUCTION

As the discussion of defining 5G evolved more than 6 years ago, the challenge was to find the differentiating properties. Clearly, the race for higher data rate would continue on, requiring an increase in data rate of 10x every five years [1,2]. For 5G this amounts to meeting a peak data rate of at least 1Gb/s at its time of introduction around 2020. But then being able to grow to 10Gb/s within 5 years, and even reach 100Gb/s around 2030, when 6G will most likely be introduced.

Many aspects of improving 4G can be collected into a common 5G vision, as has been done manifold. One good first example is given in [3].

However, it must be realized that 5G is not only about developing a set of incremental improvements over 4G, mainly in terms of increasing the achievable data rate. In stark contrast to previous improvement steps when going from 2G to 3G to 4G, the step to 5G now also opens complete new application domains, specifically within IoT and the Tactile Internet. This results in a plurality of new research challenges for building networks, as well as for developing the hardware/software architecture to efficiently support the wide spread of demand in its implementation.

This paper first sketches the dimensions of 5G in its application domains and requirements. Following a brief analysis of the signal processing and edge computing requirements is outlined, clearly showing that very different solutions are required to address the hardware/software implementation requirements.

II. THE MASSIVE DATA RATE

Memory sizes continue to increase along Moore’s Law, not only due to semiconductor scaling, but more recently also due to the introduction of 3D chip stacking. The larger embedded memories become, the larger is the size of data which needs to be exchanged. It therefore is natural that the data rate of the air interface has been growing along the same line.

For the next 100x data rate increase which needs to be challenged by 5G, multiple

- Increase in available bandwidth, also by addressing new frequency bands.
- Massive MIMO, increasing the SINR per channel, the spectrum re-use within cells, and the number of MIMO streams addressable per link.
- Increase in SINR by interference cancellation and interference coordination (ICIC and CoMP: coordinated multipoint [4]).
- Increase in spectral efficiency, in particular by filling-out the spectrum mask more efficiently and therefore reducing guard bands. This is also a requirement as carrier aggregation of non-contiguous bands becomes prevalent. Approached being discussed are multi-carrier modulation with filtered sub-bands and/or sub-carriers, as e.g. GFDM [5,6] and others e.g. as given in [7].
- Millimeter wave solutions for addressing massively available spectrum [8,9]

We can expect that combinations of the above technologies will make it possible to continue “the wireless roadmap” [1,2], i.e. increasing the maximum available data rate by 10x every 5 years. At least for addressing the needs of the 5G time window, from 2020-2030.

III. THE MASSIVE CONNECTIVITY OF IoT

With the introduction on the NB-IoT standard (narrow-band IoT) into the 3GPP roadmap in June 2016 [10], we will soon have the capability to connect 50000 or more devices per
base station cell, each connected at a very small data rate. This way a multitude of sensors will be able to be connected, sending small packets of data at intermittent time instances. Clearly, with this amount of sensors connected, one main boundary condition during the design of the NB-IoT standard was to meet a price corridor that allows for "terminals" at the US$/EUR 1 price range.

With this in mind, the application domains for this technology will be:

- Gathering information from sensors deployed widely, as e.g. parking lots, temperature, humidity, buildings, and more. It easily can be calculated that the market for this will be on the range of at least 10 billion units per year.

- Due to cellular communications having a back channel, NB-IoT will also allow for switching devices on/off. Hence, turning light bulbs, sprinkling systems, using it as a wake-up radio, and many more applications can be foreseen. In addition, it is ideal for delivering information to alphanumeric displays, as e.g. price tags in retail stores. In summary, this is a market of possibly 100 billion units per year.

- If NB-IoT is combined with positioning, it will be the infrastructure for (globally) tracking devices. Not only the lost keys will be found, also all packages in delivery can be tracked in supply chains and logistics channels, as well as retail goods in stores. The market opportunity for tracking devices can lead far beyond 100 billion possibly reaching 1 trillion units per year.

The NB-IoT standard developed so far does not yet allow for devices to transmit a packet every 100s over a 10-year duration off a AAA battery (1000mAh). Also, positioning can clearly be improved by changing the pilot signals as well as combining it with Bluetooth for the enhancement of local accuracy.

Hence, we will see that NB-IoT is a “pre-5G” standard, enabling a whole new set of applications. With its much improved version to be defined within 5G, a whole new set of applications will evolve. The wireless network will become the ubiquitous network for accessing and locating embedded sensors world-wide.

As can be seen by the discussion above, the hardware/software architectures defined for handsets today are clearly not fit for this huge field of applications, neither in power consumption nor in cost.

IV. THE TACTILE INTERNET

The third and new domain which is addressed by 5G and not by any other global standard to date is the Tactile Internet. In the previous two sections the wireless infrastructure for the delivery of content was covered. However, so far controlling the motion of real and virtual objects is only carried out via point-2-point remote control systems, and not yet by making use of a ubiquitous infrastructure. The main reason for this is that the latency of networks does not match the 1-10ms requirement. The Tactile Internet is an infrastructure for enabling remote control via the network [11,12].

Application areas can be classified and divided between being mission critical and non-critical. If an automated/remote driving infrastructure controls road intersections without traffic lights, obviously the reliability and availability of the Tactile Internet infrastructure must be high. Similar arguments hold for wirelessly controlled robotics in automated factories of the future. A loss of packets must be lower or equal 10^{-5}.

In case of gaming or education examples, however, packet losses in the order of 10^{-2} seem to be acceptable. Since this application area most likely will be the first to enter market, the first focus of research can be on meeting the latency target, adding reliability and survivability, i.e. resilience, in a later stage.

At this point the IEEE 1918.1 Tactile Internet standardization committee is in operations, building the framework and requirement specifications for the Tactile Internet [13].

The main implementation challenge of the Tactile Internet is to meet the end-to-end latency of 1-10ms. This latency budget includes the complete chain of:

- Sensor signal capturing.
- Sensor embedded system processing including the real-time operation system.
- Wireless communications to the network.
- Network delay to the control computer.
- Control processing.
- Network delay to the radio network.
- Wireless communications to the nodes being controlled.
- Embedded signal processing including real-time operating system/ the controlled nodes.
- Signal-to-actuator and actuation delay.

The easiest way to minimize the latency challenge is to position the control processing into the edge of the wireless radio access network. Hence, it has become natural to name this control processor the "Edge Cloud".

The Tactile Internet will enable complete new markets. It is the step from today’s point-to-point remote controls to an infrastructure based remote control network, the Tactile Internet, to remote control real and virtual objects. It can be seen as a similar large enabling step as was the step from cordless telephony to cellular telephony, also a point-to-point system being replaced by a ubiquitously available cellular system. The research challenges are large, and the impact to the economies will be enormously large.

V. THE PROCESSING PLATFORM – BASE STATION

To minimize the latency of communication links, in its advanced case, the edge cloud is merged with the base station signal processing into one hardware unit. In the extreme case, even this merging happens on the Multi-Processor System-on-
Chip (MPSoC). A first example for this is given in the Tomahawk-2 [14].

5G mobile communication systems will face conflicting design constraints. On the one hand, the expanding variety of transmission modes calls for highly flexible solutions supporting the ever-growing number and diversity of application requirements. On the other hand, stringent power restrictions covering from base stations merged with edge cloud servers all the way to IoT femto base stations must be considered, while satisfying the demanding performance requirements.

In order to cope with these issues, existing software defined radio (SDR) platforms, e.g. [15,16], propose an MPSoC with a heterogeneous array of processing elements. MPSoC solutions provide programmability and parallelism yielding flexibility, processing performance and power efficiency.

To schedule the resources and to apply power gating, a static scheduling approach is typically employed in today’s solutions, e.g. [15,16]. In contrast, the Tomahawk platform proposes a heterogeneous MPSoC platform (e.g. Tomahawk2 [14]) with runtime scheduling and fine-grained hierarchical power management. This solution can fully adapt to the dynamically varying workload and semi-deterministic behavior in modern concurrent wireless applications.

The proposed dynamic scheduler, named CoreManager, can be implemented either in software on a general-purpose processor or on a dedicated application-specific hardware unit. It is evident that the software approach offers the highest degree of flexibility; however, it may become a performance-bottleneck for complex applications. A first high-throughput ASIC implementation of the CoreManager was presented in [17], but this solution does not permit scheduling algorithms to be adjusted. In contrast these limitations have been overcome by implementing the CoreManager on an ASIP (application specific instruction set processor). For further details on the Tomahawk concept see [18].

Summarizing, combined base station and edge computing of the future must be flexible at many levels, allowing for software portability independent of the processor parallelism of a hardware realization. It must be flexibly adapting to the current work load of the system and the scheduling must map onto the available hardware resources to match differing optimization goals, as e.g. minimal power or minimal processing latency. The Tomahawk platform is a first step in this direction, with further innovations to follow.

VI. THE PROCESSING PLATFORM - TERMINAL

In today’s world of 4G, a diversity of terminals exists ranging from phones to embedded data terminals. The diversity is large enough to allow multiple different chip-sets to be available, tailored for the different application domains.

However, the diversity in the context of what will be seen within 5G is far too large to be able to allow for a few chip-sets to address the different needs. Does this automatically lead to a fragmentation? For sure this does lead to many new opportunities for creating fab-less chip-set startups, as the many market segments will be too small initially for large corporations to enter these with dedicated solutions.

Later on, as a second consequence it also might become a driver for consolidation at the application level. Solution companies with a diverse portfolio of 5G needs will not want to have to carry the load of supporting too many the software development platforms next to each other.

A way forward is to find hardware/software platforms that scale. They must allow for scaling along multiple requirements:

- Extreme low-power hardware solution, including voltage, clock and frequency scaling.
- Energy adaptive scaling according to the current need of signal processing.
- Hardware/software architecture scaling from small to large system implementations, including the memory subsystems.
- Software adaptive scaling and re-mapping to allow for built-in resilience as hardware units can fail temporarily in future semiconductor technologies (soft errors).
- Automatic remapping of source code and scaling according to the size and feature set of target hardware platform.
- Parallel processing scaling for scale-up and scale-out approaches of many-core platforms.

The Tomahawk platform mentioned in the previous section is clearly one approach where these scaling challenges are being researched and addressed. Embedding this vision into larger systems, the HAEC idea comes in handy [19].

ACKNOWLEDGMENT

The author wants to explicitly acknowledge the work of his entire team at the Vodafone Chair for many contributions and discussions which have led to the vision outlined here. Also, to Vodafone for its continued support and funding of the Vodafone Chair activities, in particular for Luke Ibbetson and Anna Dimitrova and their teams. Also acknowledgement and special thanks shall be given for all the fruitful scientific discussions with colleagues within the different research centers, in particular cfaed, HAEC, and the 5G Lab Germany. In the latter case special thanks to my colleague Frank Fitzek, who joined TU Dresden recently and has made discussions very inspirational.

REFERENCES
