Challenges and Solutions in 5G Infrastructure Construction

Author: Haocheng Lin

Master’s thesis in computer engineering

Organizations:
Åbo Akademi University, Faculty of Science and Engineering, Information Technologies
NOKIA S-BELL

Date: 03/02/2021

Supervisors:
Sébastien Lafond from Åbo Akademi University
Kefeng Liu from NOKIA S-BELL

Abstract

5G, as a new telecommunication generation, is growing rapidly. Meanwhile, problems occur. All previous telecommunication technology generations (from 1G to 4G) connect humans, but 5G aims at connecting objects. 5G has three axes: eMBB (enhanced Mobile Broadband), URLLC (Ultra Reliable Low Latency Communications) and mMTC (massive Machine Type Communications). Each of them has related specific use cases. The strict requirements posed by these targets imply high difficulty in 5G infrastructure deployment which will be discussed in this thesis. Besides, limited by the laws of physics, high-frequency waves used by 5G have smaller covering radii, which necessitate more antennas and more cells that lead to higher energy consumption. Energy-saving solutions are provided in this thesis for both hardware and software sides.

Key Words

Telecommunication
5G Infrastructure
5G & IoT
Power Saving

Preface

As the deployment of 5G started globally recently, the 5G topic became one of the most popular topics at the end of the decade. I, as a computer engineering student who have also studied telecommunication before, am interested in this topic. I would like to understand what the challenges in this industry are and how these problems can be overcome.

I would like to thank Nokia, one of the leaders in the telecommunication industry, which offers me an internship opportunity which allows me to gain some insights on this topic. I would also like to thank Mr. Kefeng Liu from Nokia and Mr. Sébastien Lafond from Abo Akademi who supported my thesis work. Besides, Ms. Iris Lindahl-Raittila helped me improve the quality of the language. This thesis could not have been done without their help.

Table of contents

Chapter 1

Introduction

5G technology is one of the most popular concepts in today’s world. Compared to 4G, 5G uses a higher-frequency band wave. As a result, 5G provides lower latency and higher transmission speed which can reach 1GB/s. Once 5G penetrates our daily lives, our ways of productive cooperation may change dramatically. Thanks to the low latency and high transmission speed, more and more computations can be done in the cloud, so the computing capacity of our personal devices will no longer be that important. For example, playing big games on mobiles can be possible. This may change the whole game industry. The same logic also applies in other industries. However, nothing is given without a disadvantage in it. Even in the initial infrastructure construction phase, 5G already faces many challenges.

The first are physical limitations. Higher-frequency band wave provides higher speed, but also limits its diffraction. It means it is more difficult for 5G signals to pass through obstacles. Meanwhile, 5G signals attenuate fast in the air. These two facts imply that to cover the same area, more sites are needed for 5G than for 4G. As a result, the installation and maintenance of these sites will be a huge burden for operators. More sites also mean more electricity consumption. Although 5G has way better cost performance than 4G considering the electricity consumed per bit, 5G sites consume more electricity than 4G sites since many of them are running with a low load which leads to waste. Before 5G, electric charge covers in average 20% of operators’ OPEX. If 5G sites work in the same way as 4G sites, which means they are always on, the electricity consumption will become at least three times higher than before. This is unacceptable for operators.

Thus, saving power is an important subject for operators.

The second are high-end use cases. 5G aims at three subjects: eMBB (enhanced Mobile Broadband), URLLC (Ultra Reliable Low Latency Communications) and mMTC (massive Machine Type Communications). Each of them refers to multiple special domains such as autonomous driving, high definition live and IoT. The requirements of these domains are strict. 5G should not only be an enhanced version of 4G but needs to be a revolution upgrade. 5G needs to improve the performance without sacrificing the reliability, thus, there are many technical challenges to overcome.

The third is hardware deployment. As indicated above, a large number of new sites need to be deployed to ensure 5G services. 5G cannot totally replace previous generations of telecommunication technology soon, thus, it will coexist with 4G, 3G and even 2G for a long time. Siting resources are scarce and precious and are mostly occupied by already built stations. Thus, for macro sites, upgrading existing stations by combining 5G modules is necessary, but then the limitation of space and energy supply becomes a problem. In addition, because of the small coverage radius of 5G, many micro sites need to be deployed as well, which requires solutions for networking, site selection and energy supply.

The last are social factors. Unlike enterprises, governments are born to be risk averters. This is determined by governments' functions and organization structure, so we should not blame the governments for their conservation. Therefore, governments are nearly always cautious of new technologies, relevant policies are usually introduced tardily after long periods of investigation and discussion, which objectively retards the deployment of 5G. To accelerate de deployment of 5G, enterprises need to supervise and urge the government to make laws, policies and regulations. Besides, quite a few people, especially those of low education level, are opposed to the construction of base stations. Although there is no evidence that the radiation of sites is harmful, they insist that the sites should not be built near where they live and work, regardless of the fact that this will imply worse signal reception and more intense radiation released by their phones which will be more harmful assuming that their worries were true. The danger of nuclear radiation makes the public overrate the danger of radiation in their daily lives and, as a result, efforts should be made to either change the public’s mind or conceal the base stations. It seems to be unnecessary in a technical sense but it is a real problem to be solved.

In this thesis, we will discuss the challenges above and possible solutions. We will first go through the evolution of telecommunication to obtain an overview of 5G. Then, in the theory part, we will discuss technical challenges including physical limitations, use cases, indoor coverage and site disguise. Lastly, in the practical part, a power-saving strategy for base stations will be introduced in detail.

Chapter 2

Evolution of Telecommunication

In this chapter, we will discuss some technical issues in the information transmission domain. We will review the development of telecom industry and analyze why 5G is so different.

2.1 History of Telecommunication

The world's development is highly dependent on the revolution of communication mode. The transmission of information requires a carrier. In ancient times, the carrier was limited to sound and visual signal. The longest distance for information transmission was short. Efforts were then made to prolonge this distance.

Information transfer is always important, especially in wars. Smoke towers were built to report the state of the enemy ---- visual signals were used; War drums were used to command the army ---- sound was used.

These tools augmented the distance and efficiency of information transfer. This was not only about amplifying signals but was also about encoding information. In the smoke tower example, the fire being on means enemies are approaching, no fire means the opposite. This actually transmits *1 bit* of information. As for the war drum, different commands are encoded into different rhythms. To assure the correct transmission of commands, redundancy is also taken into consideration. The war drum can transmit several bits of information, so multiple commands such as advance, change of formation and withdrawal can be transmitted to soldiers.

In the two examples above, visual signals can be transferred further. Ancient people may have understood this fact but did not know the essence. The factor that limits the propagation of sound is that it is a mechanical wave. The propagation of mechanical wave requires a medium while light ---- electromagnetic wave does not. This essential difference results in the fact that today's telecommunication mode, especially for long distance, is mainly dependent on electromagnetic waves.

In the 19th century, the telegram was invented. It made instant ranged communication possible. Morse code was invented at the same time. This is an advanced encoding method compared to those mentioned above. In theory, it can represent any information that we want to express. As a mature encoding system, Morse code was widely used. Thanks to its simplicity and popularity, until now, some people are still using it under certain conditions. When the telegram was just invented, the sending and receiving of information relied on cables. Information could only be transmitted between devices that were connected by wires. As a matter of course, the possibility of wireless telecommunication became a hot topic for scientists.

Radio communication was then invented around 1900. In the beginning, Morse code was still used. Then, with the development of technologies such as FM and AM, more complex signals, such as sound and image could be transmitted.

In the 1980s, 1G (first-generation mobile communication technology) appeared. This was the first time that telecommunication was implemented on a mobile device. In the following 30 years, the technology evolved fast. From 1G to the on-construction 5G today, we have made great progress on mobile telecommunication.

2.2 From 1G to 5G

1G to 5G are all based on electromagnetic wave communication. The electromagnetic wave has two main characteristics that are inversely proportional : frequency and wavelength. They satisfy the function $\lambda=v/f$ where λ is the wavelength, v the velocity of propagation of the wave which is almost the speed of light in the air and f the frequency. The higher the frequency, the more information it can contain since more waves are sent per second. Besides, the higher the frequency, the smaller the wavelength, which leads to higher attenuation rate and weaker scattering. Visible lights are relatively high frequency electromagnetic waves; it is difficult for them to pass through obstacles, so the spectrum used for telecommunication should have lower frequency than visible lights.

The first-generation mobile communication technology commenced in the 1980s. Of course, the term 1G did not exist at that time, it was not until 2G appeared that we called the telecommunication generation before that 1G. 1G was analogue, so in that era only voice communication was supported. The quality of voice signals was not satisfying and, even worse, the content could be easily wiretapped. Despite these technical limitations, there was not a global standard for the 1G network. As shown in Figure 2.2-1 below, there was not any standard between 1G and 2G. Japan, the Nordic countries, the U.K., and the U.S.A all had different standards, which made voice communication between different countries impossible.

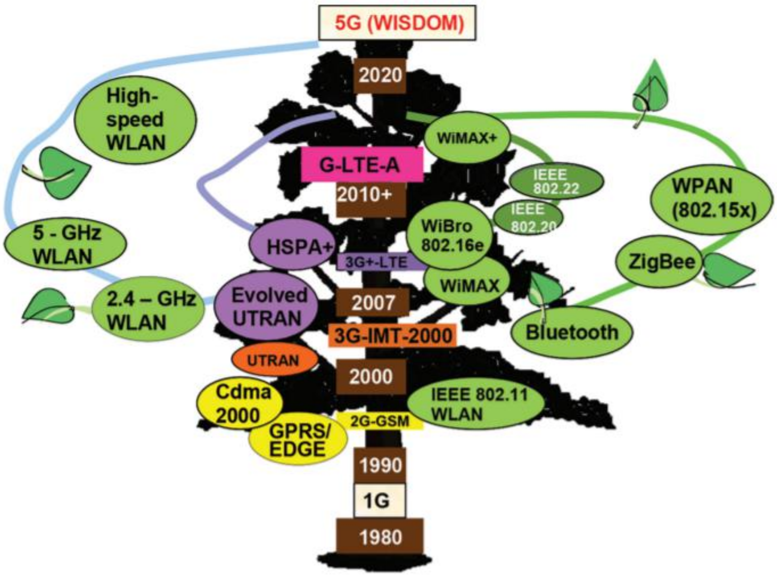


Figure 2.2-1: Tree of standards[21].

Compared to 1G, 2G had improved considerably. The signal changed from analog to digital, and the rustle in phone calls disappeared. In addition, sending verbal content (SMS) became possible. The most remarkable difference between 1G and 2G is that 2G was based on a globally agreed standard: GSM, Global System for Mobile Communications which was set in 1991. However, the maximum bandwidth was still quite limited, only 9.6kbps. In the hope of transmitting more complex data_such as images and videos at an acceptable rate, a revolution was still needed.

3G was designed for high speed data rates. Although the so-called “high” speed seems ironic today, it was a great progress at that time. 3G first connected mobile devices and Internet and provided many other useful services such as video calls and GPS. The spectrum for 3G was at about 2000MHz while that for 2G was at about 900MHz.

4G, which we use the most today, offers a very high data rate. In an outdoor environment, while 3G provides usually several hundred kbps, we can easily access several Mbps with 4G. The transmission speed increases while the tariff decreases; thus, watching videos online using mobiles is no longer an extravagant action. Thanks to 4G, we can now view web pages, watch videos or live without spending too much time on loading.

5G aims at high data rate, low latency and more devices. For data rate, the unit will evolve from Mbps to Gbps. At this data rate, the transmission time for any common sources can be ignored. When we download a file, the “downloading” stated will disappear, we will see that files are fully downloaded instantly. To support this data rate, the frequency used will be extremely high. The main spectrum for 5G is at about 3500MHz. There is also another frequency range: 24GHz to 52GHz allocated for 5G. The theoretical 20 Gbps data rate is calculated from this range. Using the formula mentioned in the beginning, electromagnetic waves at this frequency have a wavelength of several millimeters, so we call them millimeter waves. There are some problems to solve if we want to transmit data using millimeter waves compared to the waves with bigger wavelength that we used before. These will be discussed in the next chapter.

Chapter 3

5G Technical Challenges

In this chapter, the following question will be discussed: why is 5G so different? In the last 40 years, we have experienced four different generations of mobile communication and have experienced generation changes every decade. But the changes were smooth. We did not hear many problems that operators met in new infrastructure construction process, and the new generation topic was never as hot as today. Although the development of media makes us be more aware of the news, 5G must also have its particularities which make it different from the previous generations. These particularities will be discussed in this chapter.

3.1 Physical Limitations

The transmission of electromagnetic waves does not require a medium, it also means that in an ideal vacuum environment, the transmission of electromagnetic waves does not have any attenuation. However, the attenuation exists in the air. According to the book *Radio Monitoring: Problems, Methods and Equipment*, the losses L in free spaces can be calculated by the formula below [23]:

$L = 32.45 + 20\log(f) + 20\log(d)$

in which:

L = losses in dB
f = frequency in MHz
d = distance in km

The main spectrum for 5G macro sites is about 3.5GHz which is about 1.35 times the frequency of 4G(2.6GHz). According to the formula, we can calculate that coverage radius of a 5G site will be 1.35 times smaller than that of a 4G site. For higher frequencies, the coverage radius will continue to decrease. One 5G site may only be able to cover a radius of 100 meters. Thus, the number of 5G sites will be extremely large.

3.2 Use Cases

Each time the figure before G changes indicates qualitative change. With the change of generation, new use cases are invented. 1G brought voice chat, 2G brought SMS, 3G brought images and 4G brought videos. We can connect our mobile devices to Internet since 3G. 4G brought higher speed and lower latency, and 5G will bring ultra-high bandwidth and ultra-low latency. How will this “ultra” change our lives?

In short, 1G to 4G connect humans and 5G connects objects. In the visible future, the most important applications of 5G will be IoT and autonomous driving. Humans are usually not sensitive to the latency except in certain domains such as e-sports. While watching live or having a video chat, we can totally tolerate a 1-second delay. On the contrary, synchronization and real-time monitoring are important for machines. For autonomous driving, 100-ms latency can be fatal. The difference in the subject that 4G and 5G serve implies that despite high data rate and low latency, 5G should also be ultra-reliable.

3.2.1 5G and Autonomous Driving

Today, nearly all automobile manufacturers are developing or sponsoring autonomous driving technology. Meanwhile, many tech magnates have also invested considerably in this domain. According to the SAE standard *J3016_201609* updated in 2016, the levels of automation are from 0 to 5. Now, the commercial vehicle type that meets the highest-level requirements is *Audi A8* which is assessed as level 3: Conditional Automation. The definition of this level is *“the driving mode-specific performance by an automated driving system of all aspects of the dynamic driving task with the expectation that the human driver will respond appropriately to a request to intervene”.* [SAE STANDARD J3016] Audi A8 claims to be able to control the car safely in cities at a speed below 60 km/h. This is so far the only commercial type that reaches level 3, in addition, there is still a huge gap between level 3 and how the public understands autonomous driving, which means the system can control the car in all daily driving conditions including highways, mountain roads, country roads etc.. Many media report that 5G will be a key technology for autonomous driving. This point of view will be discussed in the following sections. We will go through the core technologies of autonomous driving and discuss for which of them 5G can be an ideal solution.

3.2.1.1 Capturing Surroundings Information.

Autonomous cars are equipped with sensors (as shown in Figure 3.2.1.1-1) to capture surroundings information. This information is necessary for them to interact correctly with the environment. In today’s situation, not many **digital** data can be collected while driving, the autonomous driving is highly dependent on the information captured by sensors. In this area, 5G has no impact.

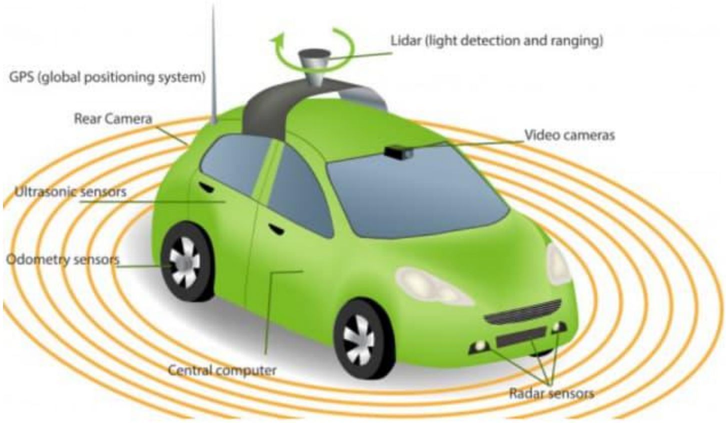


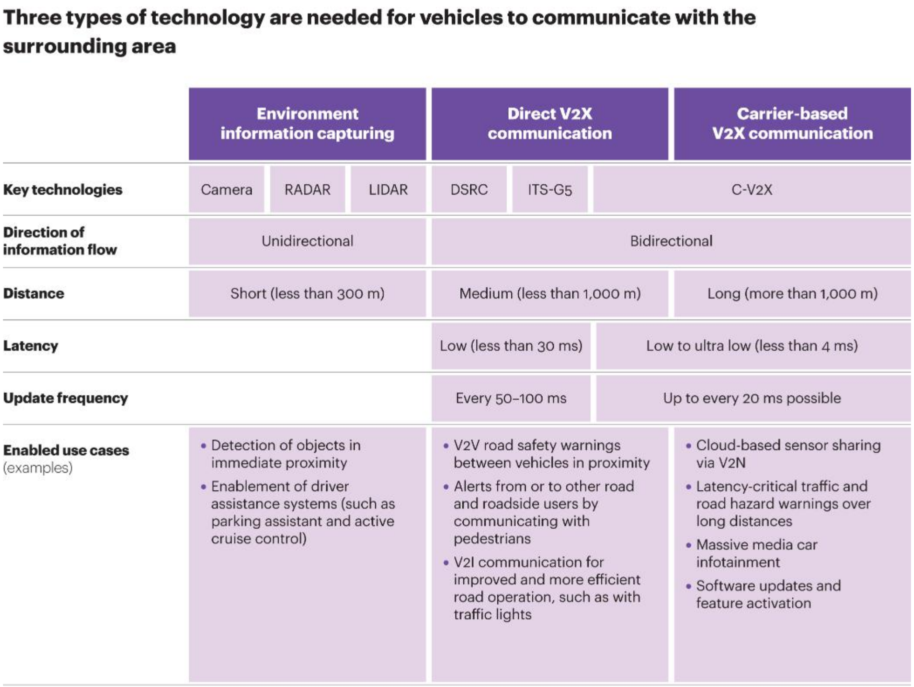
Figure 3.2.2.1-1: Autonomous car with sensors.

[https://www.engineering.com/story/18809]

3.2.1.2 V2X Communication

V2X is the abbreviation for vehicle to everything. Sensors are not perfect, furthermore, analyzing the information captured is also a big issue. Thus, the vehicle needs to connect with all important objects nearby, including traffic infrastructure and other vehicles. Now, the 5.9GHz spectrum is dedicated to this functionality. Today, "*Two V2X communication approaches are being pursued*", one of them is based on Wi-Fi and the other is based on 5G standards. [14]

As we are looking forward to full autonomy, V2X is essential. Digital data are much easier to analyze than other forms of data collected by the sensors. If all necessary data are collected in digital form, autonomous driving is like making a racing game which we already have many of in the market. Considering the importance of V2X, setting a common standard for it is the first priority. As mentioned above, there are two approaches: US favors the standards of DSRC which are based on Wi-Fi, IEEE 802.11p and China uses Cellular-V2X which is now based on 4G LTE but on 5G in the future. Europe is now using ITS-G5 which is based on Wi-Fi but is still hesitating between these two approaches. In Figure 3.2.1.2-1 below, we can see the difference between these two approaches.


Three types of technology are needed for vehicles to communicate with the surrounding area

	Environment information capturing	Direct V2X communication	Carrier-based V2X communication
Key technologies	Camera RADAR LIDAR	DSRC ITS-G5	C-V2X
Direction of information flow	Unidirectional	Bidirectional	
Distance	Short (less than 300 m)	Medium (less than 1,000 m)	Long (more than 1,000 m)
Latency		Low (less than 30 ms)	Low to ultra low (less than 4 ms)
Update frequency		Every 50–100 ms	Up to every 20 ms possible
Enabled use cases (examples)	• Detection of objects in immediate proximity • Enablement of driver assistance systems (such as parking assistant and active cruise control)	• V2V road safety warnings between vehicles in proximity • Alerts from or to other road and roadside users by communicating with pedestrians • V2I communication for improved and more efficient road operation, such as with traffic lights	• Cloud-based sensor sharing via V2N • Latency-critical traffic and road hazard warnings over long distances • Massive media car infotainment • Software updates and feature activation

Figure 3.2.1.2-1: Technology need for autonomous cars[14].

From the figure we can see that in all three important dimensions: distance, latency and update frequency, 5G-based C-V2X is better than Wi-Fi based DSRC and ITS-G5. Thus, there must be another factor that makes Wi-Fi competitive with 5G in this domain. This factor is its maturity. Wi-Fi has evolved for more than twenty years and has been deployed all over the world, however, in the second half of 2020, the commercial deployment of 5G has just begun. Wi-Fi now supports medium range communication (from 300m to 1km) and for fast-moving objects, Wi-Fi also provides reliable services. As for the latency, Wi-Fi provides shorter than 30ms latency, which is enough since some time-critical applications only require the latency to be shorter than 100ms. With the addition of the relatively low deployment cost, Wi-Fi might be a good option to kick-start autonomous driving.

3.2.1.3 5G's Role

Many media assert that 5G is one of the key technologies for autonomous driving, however, we have seen that 5G is not involved in environment information capturing and is not the only option for V2X either. Thus, why did we always discuss 5G and autonomous driving together?

Indeed, 5G is not the only option at the beginning. However, we have discussed that V2X is essential to autonomous driving. Wi-Fi provides acceptable performance but still has some problems. The first one is region switch. Now, when we connect our devices to Wi-Fi, when the signals are weak, the devices will not automatically connect to a better Wi-Fi or use 4G to access Internet. The switch is not smooth. The distance supported by Wi-Fi is 300m to 1km. Thus, for cars, region switches will be quite frequent. Considering the bad switching performance, the quality of services cannot be guaranteed. The second problem is the maximum number of connected objects. Providing Wi-Fi services for large-scale activities such as concerts and ball matches is one of the major businesses of certain companies, but the quality of connection is not always reliable. In the future, we expect V2X to be fully implemented, which means the density of connected devices, including cars and traffic infrastructure such as traffic lights, will be huge. One advantage of 5G is that it can support big density of connected objects: up to 1 million objects per square kilometer (we will discuss this subject later in the IoT sub-section). In this sense, 5G is one of the key technologies for autonomous driving.

3.2.2 5G and IoT

The concept IoT (Internet of Things) was proposed about twenty years ago. In recent years, it was widely discussed by IT industry practitioners and common people because there was a public opinion that 5G technology would assist IoT immensely, and the deployment of 5G could be seen in the near future. However, to date, 5G has not been widely deployed yet, but there are already many IoT devices working, such as intelligent electricity/water meters and intelligent wearable devices. Thus, what is the real relation between IoT and 5G? How important is 5G's role in a "real" IoT world?

There is no clear definition of "IoT device". Thus, I will use my definition to identify IoT devices and briefly classify them into several types. The classification method may not be fully objective but can at least lead to better discussion of how 5G will influence IoT.

In my opinion, IoT device can refer to any device that can process data, transmit and receive data wirelessly, while Internet connection is not one of its main functionalities. According to this definition, any device that can connect to Internet except phone, tablet, laptop, computer can be considered as IoT device. For an IoT device, we have several indicators to evaluate its performance: data rate, energy consumption, reliability and latency. Some devices have no strict requirements on any of these four aspects (such as air conditioner that can be controlled remotely by phones), but some have critical requirements on one or more aspects. For these devices, 5G will be helpful for their performance since 3GPP (3rd Generation Partnership Project) defined three major aspects for 5G: eMBB(enhanced Mobile Broadband), URLLC(Ultra Reliable Low Latency Communications) and mMTC(massive Machine Type Communications) that correspond to different characteristics. We will see some examples.

3.2.2.1 Low Energy Consumption

Low energy consumption, wide range communication and long service life are the characteristics for many IoT devices, such as water meters and environment monitoring sensors. Besides, the number of such devices is big, which means large density of connected objects. Thus, we need an LPWAN (low power wide area network) to connect them. This is what mMTC involves. Some requirements for mMTC are listed below: [4]

- *battery life of 10 years*
- *coverage penetration of 164 dB with throughput of 160 bits per second*
- *coverage density to support of up to a million devices in a square kilometer*

NB-IoT (Narrow band IoT) is one of the solutions for this kind of IoT devices. 3GPP started to set the standards for it in 2015 and completed the standards in 2016. On July 9th, 2020, ITU (The International Telecommunication Union) recognized NB-IoT as a 5G standard[11], so it is now officially a part of 5G technology.

NB-IoT removes unnecessary protocols to reduce costs. As for the reliability, it supports re-transmission up to 128 times for the upstream. On the hardware side, NB-IoT uses a single antenna and only supports half duplex. Compared to full duplex used in 4G LTE, half duplex means there is no need to send and receive data at the same time, which will save much energy. One main difference between this kind of devices and full duplex devices such as cellphones is that these devices only **send** information most of the time and **do not need to receive** signals. Thus, DRX (Discontinuous Reception) can be updated to eDRX (extended DRX). Cellphones are not always trying to receive signals. When they are idle, they monitor possible signals fitfully to save energy(about once every three seconds). eDRX extends this interval from maximum 2.56 seconds to maximum 2.91 hours [20]. It will dramatically reduce the energy consumption of these small IoT devices. In addition, for these ultra-low power-consuming devices, technologies have been developed to harvest energy from the surrounding space. “A specialized circuit converts some of the ambient RF energy into electrical power to charge a battery.”[8] These technologies contribute to the extremely-long battery life.

3.2.2.2 High Data Rate

High data rate mainly involves high definition live stream. The deployment of 4G has made watching online videos and live from mobile phones possible. Our daily requirements for videos have been generally satisfied. However, in some areas, such as security and protection, higher definition is still required.

Today, we often discuss 4K videos. 4K mainly refers to 4096 * 2160 definition, some other scales such as 3656 × 2664 can also be considered as 4K. To transfer a 4K definition video in real time, we need originally a data rate of 4096 * 2160 * 8(8 bits/color) * 3(three colors) * 60(frames/s) ≈ 11.9Gbps. With a compression rate of 300 to 1000, we still need about 25Mbps. Although 4G can theoretically provide 100Mbps, the data rate is usually around 10Mbps in real life. On the contrary, 5G can support 4K videos much more easily.

4K is the definition of most of the movies that we watch in the cinemas today. At this definition, we can even see the fur of the characters in the movie. Figure3.2.2.2-1 below is a surveillance image of Qingdao train station in 2017. The definition of it is 720p (1280*720). At this definition, we can recognize some characteristics of passengers: their clothes, their stature, the shape and color of their belongings, their actions etc. It is already helpful for recognizing potential thieves and tracking them. However, much manpower is still needed.

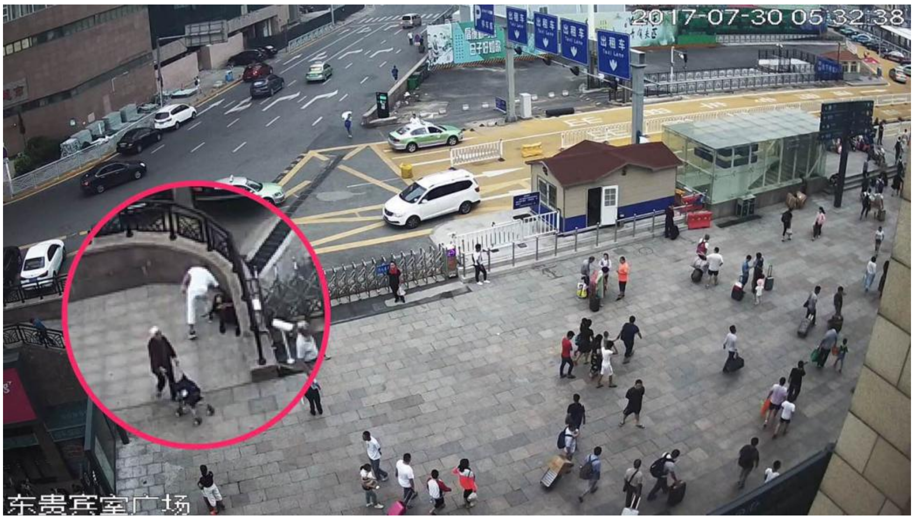


Figure 3.2.2.2-1: A surveillance image of Qingdao train station. [Sohu]

If the definition augments to 4K, much work can be handled by machines. At this definition, we can see more details, such as the facial characteristics and the scars of the passengers. By using the face recognition technology, the identification of passengers will become much easier.

Not only in this kind of crowded places will 4K definition be helpful. 5G is a wireless technology, which means the deployment of application devices is convenient. We can easily equip a drone with a 4K web camera, it can then transfer high quality live stream. Whenever an emergency occurs, the live report can be much clearer than before. Thus, 5G will be a breakthrough in the security and live domain. This is what eMBB involves.

3.2.2.3 Low Latency

Low latency is as important as high data rate in 5G. IoT has many time-critical use cases. Time-critical in most of the time also means that long latency or packet loss can cause harmful consequences, thus, these use cases also require high reliability.

In June 2020, Ericsson published an article listing the main time-critical use cases [2]:

- *Industrial control*
- *Mobility automation*
- *Remote control*
- *Real-time media*

For more detailed use cases, see Figure 3.2.2.3-1 below:

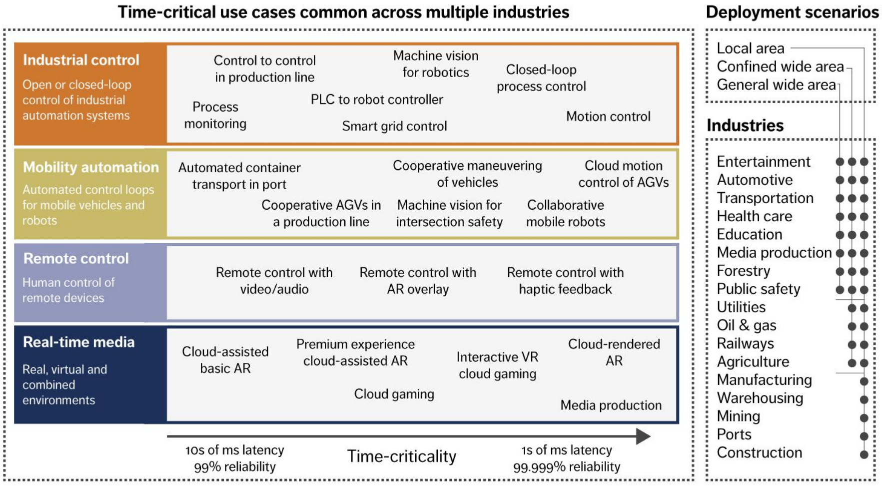


Figure 3.2.2.3-1: Examples of use cases enabled by Critical IoT. [2]

3.2.2.3.1 Demonstration of the Low Latency

On Nov 22, 2018, French football club Paris Saint-Germain uploaded a video on YouTube showing an interesting experiment of 5G latency powered by Ericsson[22]. In this experiment, players are asked to play football while wearing a special pair of glasses. The images that the players will see are the images recorded by the camera on the glasses, uploaded to the control room and downloaded to be displayed in the screen in the glasses (see Figure 3.2.2.3.1-1 below). In short, they see the images delayed for the latency of 5G.

Figure 3.2.2.3.1-1: How video signals propagate in PSG's experiment of 5G latency.

The experiment shows that in 5G condition, players' actions are correct but in 4G condition they do not manage to save, juggle or even pass the ball. Although it is obvious that some players' actions in 4G condition are exaggerated in the video, we can still trust that in 5G condition, the latency can be almost ignored even by football players. Thus, the latency is acceptable in many other time-critical use cases.

3.2.2.3.2 How the Low Latency be Implemented

When we discuss the **latency** in 5G, most often the latency refers to the air interface delay: the time needed to send and receive data from a site wirelessly. Once the data enter the core network, they are transmitted by fiber Thus the transmission speed is theoretically the velocity of light and cannot be further improved.

The latency (air interface delay) in 4G is theoretically 20ms. The uplink process takes 12.5ms and the downlink process takes 7.5ms. Figure 3.2.2.3.2-1 and 3.2.2.3.2-2 below are the signaling chart. Table 3.2.2.3.2-1 and 3.2.2.3.2-2 below show components of the delay. [12]

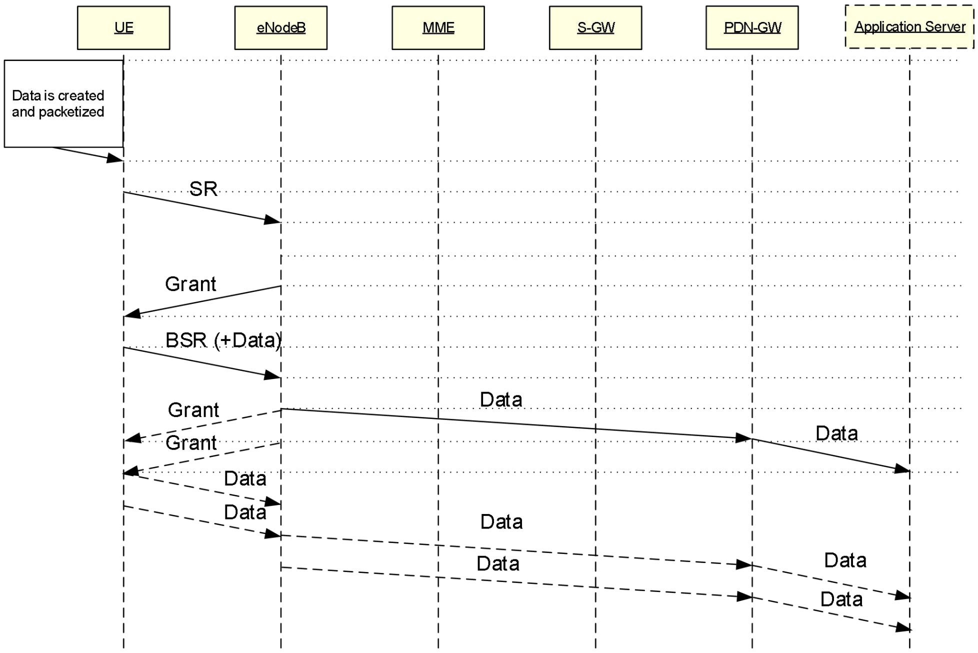


Figure 3.2.2.3.2-1: Overview of UL transmission delay. Does not include re-transmissions[12].

Table 3.2.2.3.2-1: Typical radio access latency components (Rel. 8/Rel. 9) for a UL transmission from a UE without a valid uplink grant[12]

Component	Description	Time (ms)
1	Average waiting time for PUCCH (10 ms SR period/1ms SR period)	5/0.5
2	UE sends Scheduling Request on PUCCH	1
3	eNB decodes Scheduling Request and generates the Scheduling Grant	3
4	Transmission of Scheduling Grant	1
5	UE Processing Delay (decoding of scheduling grant + L1 encoding of UL data)	3
6	Transmission of UL data	1
7	Data decoding in eNodeB	3
	Total delay [ms]	17/12.5

When an endpoint wants to upload data, it needs firstly to send a scheduling request. The site will then send a scheduling grant back, after which the endpoint can upload the data. Scheduling request, scheduling grant and data all need to be decoded. The decoding process takes 3ms each. Besides, each single-trip transmission takes 1ms. All of these contribute to the 12.5ms total delay.

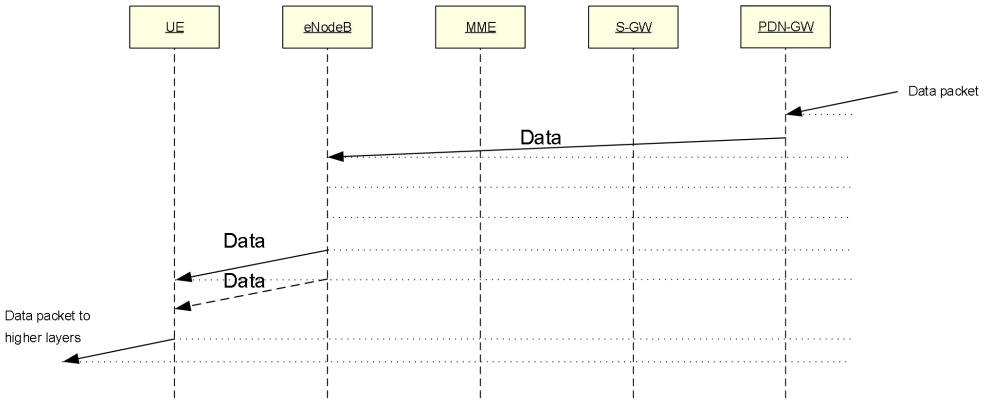


Figure 3.2.2.3.2-2: Overview of DL transmission delay. Does not include re-transmissions[12].

Table 3.2.2.3.2-2: Typical radio access latency components (Rel. 8/Rel. 9) for a DL transmission[12]

Component	Description	Time (ms)
1	processes incoming data	3
2	TTI alignment	0.5
3	Transmission of DL data	1
4	Data decoding in UE	3
	Total delay [ms]	7.5

In the downlink process, there is no scheduling grant, thus, the total delay is shorter than that for uplink.

The final target of 5G URLLC is to systematically reduce the latency to 1ms.

Improvement in 4G

As shown in the figures and tables above, 4G LTE air interface latency mainly consists of three parts: Grant acquisition, Transmission time interval and Processing (re-transmission ignored). Since the processing time is not highly relevant to wireless communication, efforts were made on the other two parts.

- Grant acquisition

When an endpoint wants to upload data, it firstly needs to demand a scheduling grant. In the total 12.5ms uplink latency, this part contributes 8.5ms. This is the time “wasted” before uploading data. To minimize this part is therefore important.

Before LTE release 14, pre-scheduling was adopted to reduce the latency. Sites would periodically allocate uploading resources to end users, so they did not need to request a scheduling grant each time they wanted to upload data. Disadvantages of this method are evident:

no matter whether the end users need to upload data or not, they are given resources to do so. This will cause waste of wireless communication resources.

The scheduling method later evolved from pre-scheduling to semi-static scheduling. The only main difference is that in semi-static scheduling, the endpoints are not forced to upload padding data when not wanting to upload data. This measure not only reduced the latency, but also reduced the energy cost of endpoints.

- Diminish transmission time interval

One LTE frame is 10ms. In the beginning, the LTE transmission interval was one subframe which is 1ms. To diminish the transmission interval, the subframe needs to be re-divided. One subframe can be divided into two slots and one slot contains seven signals. Thus, the minimal scheduling time unit was shortened to $\frac{1}{14}$ms. LTE frame structure can be seen in Figure 3.2.2.3.2-3 below.

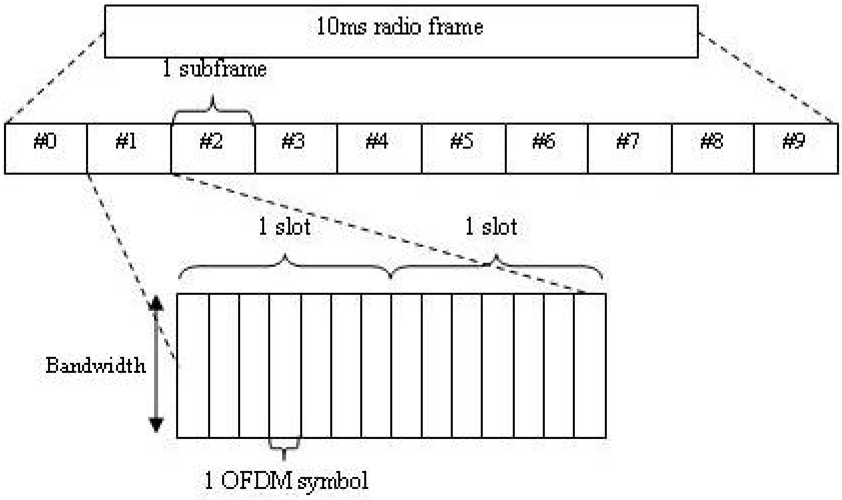


Figure 3.2.2.3.2-3: LTE frame structure[16].

These are the main measures taken by 4G to reduce the air interface latency. Meanwhile, processing time had also been reduced thanks to the augmentation of processing capability and some other factors. Until 2018, in LTE release 15, the theoretical latency had been reduced to 2.7ms. [LTE release 15]. Reducing the latency to 1ms then becomes the mission of 5G.

Improvement in 5G

- Still shorter time slot

One of the characteristics of 5G is its high frequency. It supports higher bandwidth and shorter time slot. One subframe designed for 5G is still 1ms, but the number of slots depends on the actual frequency. In 4G, the subcarrier spacing is 15kHz and one subframe contains two slots. When the subcarrier spacing is 30kHz, one subframe will contain four slots. Since the number of slots is proportional to the frequency, the shortest time unit is inversely proportional to the frequency.

- Grant free transmission

When it comes to the scheduling grant, 5G inherits the semi-static scheduling method from 4G to reduce latency. To solve the problem of wasting wireless communication resources on endpoints that do not need to upload data, 5G allows sites to allocate resources to a **group** of endpoints. In many cases, this will reduce the waste.

Low latency most of the time coexists with high reliability. Besides reducing latency, we still need to focus on the reliability. Since problems cannot be totally forecasted, the reliability is generally assured by re-transmission. The re-transmission mechanism requires the sender to track the status of the data sent which will evidently increase the complexity of the system, therefore, balancing latency and reliability is an issue for 5G.

- Async HARQ

HARQ is the abbreviation for Hybrid Automatic Repeat Request. It combines FEC (Forward Error Correction) and ARQ. HARQ was synchronous in 4G but asynchronous in 5G. Asynchronization means higher flexibility. 5G also supports preemption scheduling, which means high priority endpoints can grab the resources allocated to lower priority endpoints, thanks to what time-critical data can be processed preferentially.

3.2.2.3.3 Conclusion

The total E2E latency is eventually limited by physical rules: the speed of transmission of information cannot exceed the speed of light. In 4G, the real total latency is usually at 100ms level. It seems that the 20ms air latency is not the main part. However, when we reduce the air latency to 1ms in 5G, reducing the other parts of latency becomes significant. We can build a more well-structured content distributed system, push content to edge and eventually reduce the total E2E latency to below 10ms.

A blink of an eye takes in average 100 to 400ms. The interval between (the eyes) capturing a visual signal and (the brain) seeing the image is also approximately 100ms. Therefore, 10ms E2E delay is short enough and negligible for human beings. The improvement is thus significant.

3.3 Indoor Coverage

5G signals attenuate fast in the air and can be easily blocked by obstacles. Thus, outdoor sites cannot cover the indoor environment. Meanwhile, researches show that in 4G era, 80% of the uses happen indoors and in 5G era, this figure might augment to 90%. Today, 60% of 4G users are unsatisfied with the indoor coverage. Improve indoor coverage is an inevitable issue for the 5G deployment.

Ericsson proposes an indoor 5G deployment approach in its white paper *Bringing 5G networks indoors*[7]. Figure 3.3-1 below shows how it works in a standard enterprise environment.

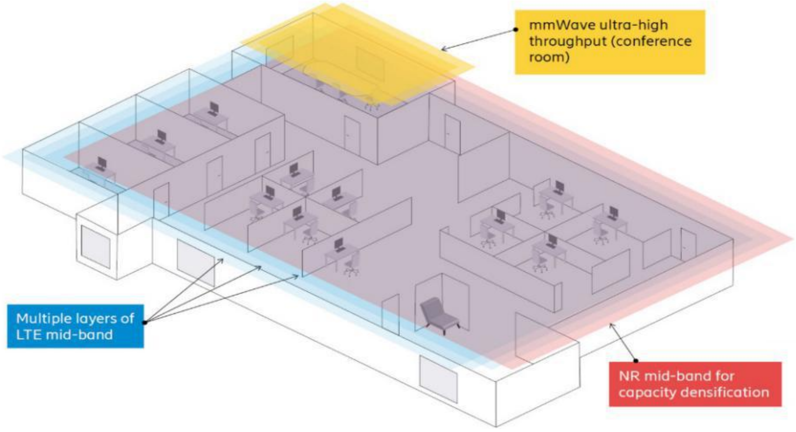


Figure 3.3-1: A deep layer approach in a standard enterprise environment. [7]

As shown in the figure, only the conference room is covered by mmWave which provides ultra-high throughput, other areas are only covered by 4G LTE. This figure demonstrates how Ericsson considers 5G: 5G will not replace 4G. They will coexist and complement each other. 4G LTE meets most of our daily requirements so it will still work for a long period of time. 5G mmWave has severe physical limitations and is not easy to be deployed in the complex environment, thus it should only be used in special places where ultra-high data rate or ultra-low latency is required.

3.4 Site Disguise

The traditional macro site, as shown in Figure 3.4-1 below, is usually a huge steel tower and is conspicuous. Because of some people's stereotype that the radiation released by sites is harmful, this kind of site is sometimes boycotted by the public. In the 5G deployment, millions of new sites need to be deployed, so the public's reaction must be considered although it is often irrational.

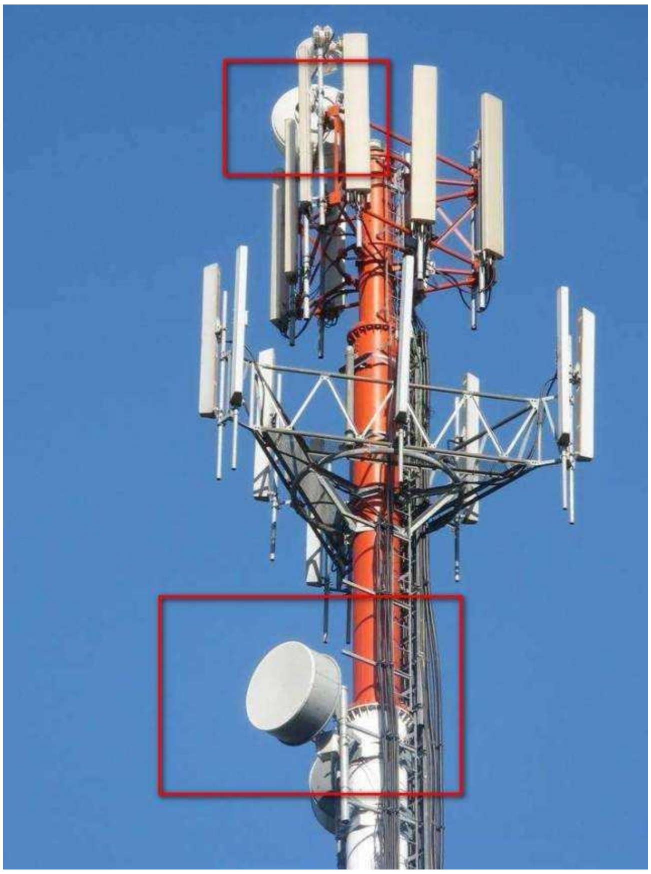


Figure 3.4-1: Traditional macro site. [sohu]

Figures below (Figure 3.4-2,3.4-3,3.4-4) are some disguised sites. Sites can be disguised as plants or streetlamp outside, and can be disguised as air conditioner external unit on rooftops. The main idea is to disguise the sites as common objects in our daily lives. In addition to avoid the boycott, it also makes the city prettier.

Figure 3.4-2: Tree-like site. [baidu]

Figure 3.4-3: Streetlamp-like site. [qdsezm]

Figure 3.4-4: Air conditioner external unit-like site. [21ic]

3.5 Conclusion

High frequency is one of the characteristics of 5G. When we discuss the challenges of 5G, worse diffraction and faster attenuation caused by higher frequency are the two main physical limitations. These two factors make 5G coverage harder for both outdoor and indoor environment.

5G has three main targets: eMBB (enhanced Mobile Broadband), URLLC (Ultra Reliable Low Latency Communications) and mMTC (massive Machine Type Communications). mMTC does not require high data rate thus the coverage problem does not affect it. URLLC involves time-critical use cases including autonomous-driving. The requirement of high reliability necessitates redundancy in the sites' deployment. eMBB is the target that is highly relevant to the high data rate, thus it is affected by the coverage problem. Since most of the uses happen indoors, indoor 5G coverage is inevitable. For this, 5G is not a superior alternative of 4G, on the contrary, 5G and 4G will complement each other. Only the hottest spots need to and should be covered by 5G signals.

Besides technical problems, social problems also need to be taken into consideration. Site disguise strategy is an important issue for equipment manufacturers. Although disguising sites does not improve their performance, it is beneficial for long-term development of the whole society.

Chapter 4

Power Saving

As indicated in chapter one, power saving is a big issue in 5G infrastructure construction. Electric charge has already been an important part of operators' OPEX, if it continues to increase rapidly, it will become a huge burden for operators. Thus, this is an important item in operators' control of expenditure. Solutions for reducing base stations' power consumption will be introduced in this chapter.

4.1 Hardware

4.1.1 Cooling

The cooling system is one of the most important components for a base station. Although the concrete data vary from different sources, nearly all sources indicate that cooling is the second most power-consuming module of a base station and it consumes about 30% of the electricity. Current cooling systems usually use a traditional air conditioner which is not really efficient. Therefore, updating the cooling system is a rational direction of reducing the total energy consumption.

Nokia Liquid Cooling Base Station

In 2018, the world's first liquid cooling base station started to operate in Finland for Elisa Oyj [18]. On June 3rd, 2020, Nokia announced that its liquid cooling stations had helped Elisa save about 30% of their base stations' energy expenses. Meanwhile, they had also reduced the CO2 emission by about 80%[19]. Figure 4.1.1-1 is a screen shot of the Nokia's promotional video.

Figure 4.1.1-1: Nokia's liquid cooling base station[18].

The base station has thick cubes of two colors. Blue tubes are for cold water and red tubes are for hot water. Main modules of the base station, including AAU and DU, are made compatible for liquid cooling as shown in Figure 4.1.1-2 below. The module has a special pipe for water circulation which allows cold water to absorb the heat generated by the module[25]. Since water's volumetric heat capacity is much higher than that of air, liquid cooling is much more efficient than traditional air cooling system.

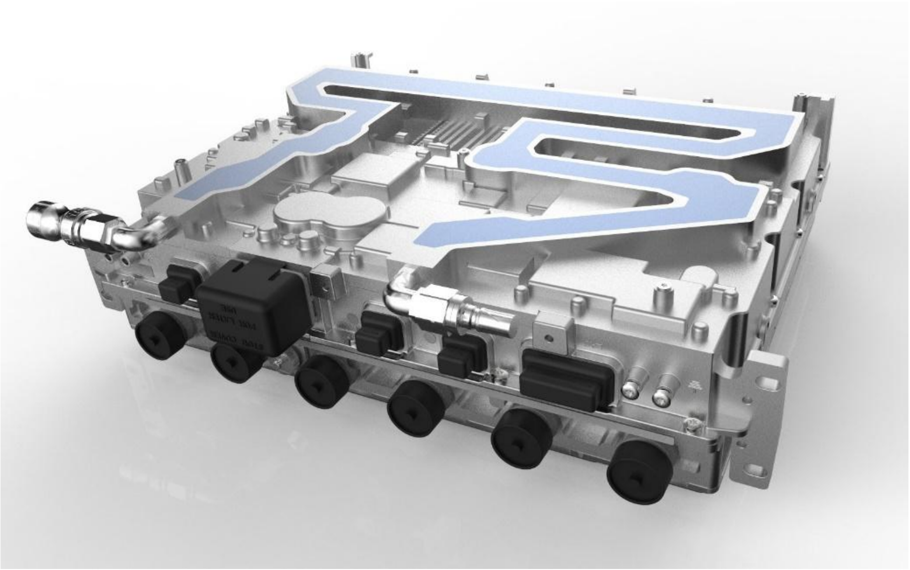


Figure 4.1.1-2: Nokia's liquid cooling module[25].

In addition to higher volumetric heat capacity, liquid cooling has another advantage that the heat collected can be easily reused. In cold countries such as Finland, heating is essential. If the site is built on top of a building, hot water can be easily reused to support the building's heating system. The reuse of heat generated helps the site reduce the CO_2 emission by 80% while saving 30% energy.

4.1.2 Power supply

Except improving the cooling system, another approach of saving energy is adopting a better power supply method.

High QoS of base stations relies on sustained and steady energy supply. In addition to the urban electricity supply (110V/220V AC in most countries), batteries are also required to cope with emergency circumstances. In the previous decade, the VRLA (valve regulated lead–acid) battery was most commonly used. As from 4G to 5G, a base station's power rate triples from about 900W to approximately 2700W, to ensure four-hour continuous energy supply in emergency situations, the battery capacity also needs to be tripled. Since space in sites is limited, making the battery three times as big as before is not a good solution.

Fortunately, together with the development of smartphone, electric bicycle and mobile, the technology of lithium battery has been much developed. The comparison of these two types of battery can be seen in Table 4.1.2-1 below.

Table 4.1.2-1: Comparison between VRLA and lithium battery

[Source: GGII]

Type	VRLA	Lithium
Service life	8 years	10 years
Charge-discharge	1000 times	10000 times
Power density	40Wh/kg	180Wh/kg
Price	0.6RMB/Wh	0.8RMB/Wh
Recycle	More than 90%	No recycle

The most interesting superiority of lithium battery is its power density: more than four times higher than that of VRLA battery. It implies that although the requirement of battery capacity needs to be tripled, by replacing traditional VRLA batteries by lithium batteries, the power supply module can become even smaller than before, which ameliorates the limitation of space.

Considering the price, the price per year seems to be similar. The 90% recycle rate of VRLA battery only involves the recycle of the material, so it only has a little impact on the price. Real situation is different from the theory though. According to Juda -- a Chinese lithium battery manufacturer -- some characteristics differ from the theoretical value when involving the power supply of base stations. The lithium battery has been tested on a real outdoor base station. A 150Ah lithium battery was used to replace a traditional 200Ah VRLA battery (the requirement is 100Ah). The lithium battery costed 11400 RMB while the traditional VRLA battery costed 5040 RMB but had only 1/3 of lithium battery's lifetime[13]. According to these data, the lithium battery was already cheaper than the VRLA battery. In addition, since the recycle of VRLA batteries is mature, there is great risk that the VRLA batteries be stolen. Although high recycle rate is a good characteristic, it augments objectively the cost of a base station.

In short, the lithium battery is almost in all aspects better than the traditional VRLA battery.

Huawei Blade Site

In 2018, Huawei released its Blade Site which got the first prize of China's National Science and Technology Progress Award. The principle of the Blade Site is that all components are "bladed" which means different modules can be combined and deployed easily and seamlessly, like Lego bricks (see Figure 4.1.2-1 below). Another feature of Huawei Blade Site is that it is self-cooling and can bear up to 55 degree Celsius. Thanks to this feature, Huawei Blade Site does not require additional cooling system and reduces directly about 40% of the total base station's power consumption.

Figure 4.1.2-1: Huawei Blade Site components. [huawei.com]

For the Blade battery specifically, the high temperature resistance relies on the lithium battery. Although traditional VRLA battery can work in high temperature environment, its lifetime and capacity will be influenced. Its standard working temperature is 20-25 degree Celsius. On the contrary, the lithium battery can work at high temperatures without sacrificing its performance.

In Table 4.1.2-1 above, we can see that besides power density, another aspect in which lithium battery performs much better than VRLA battery is the number of times that the battery can be charged and discharged. The lithium battery’s this index is also stable, not sensitive to the temperature, which implies that the functionality of the battery can evolve from **backup** power supply to **secondary** power supply. To be more concrete, since the lithium battery can be charged and discharged 10,000 times, we can charge it every day at midnight and discharge it at midday. In this case, we only need to charge and discharge it 3650 times in 10 years, it is still far from the upper limit. By doing so, we indirectly save energy since electricity at midnight is less costly than at peak time, although there will be some waste while converting energy from one form to another.

4.1.3 Conclusion

The cooling system is one of the most power-consuming systems in a base station, second only to main equipment’s power consumption. Nokia’s liquid cooling system changes the underlying logic of how cooling system works. The efficiency of liquid cooling and air cooling is at different level.

Huawei finds another approach. Instead of reducing the energy consumption of the cooling system, Huawei finds a solution to remove the whole cooling system. The reason that traditional base stations need a cooling system is that the base station equipment, especially the battery, generates considerable heat but cannot bear high temperature. In summer, the temperature inside a base station can easily reach more than fifty degree Celsius, which is unbearable for many components. By replacing the VRLA battery by the new lithium battery, the necessity of the cooling system descends. After solving this main problem, Huawei successfully made the whole base station resistant to high temperature, which augments the flexibility of deploying base stations. Besides, the use of lithium battery also implies changes in the power supply logic.

In conclusion, on the hardware side, improvements in these two dimensions will significantly reduce the power consumption of base station.

4.2 Software

4.2.1 Problem Description

In 2G/3G/4G era, a site consists mainly of three components: A BBU (building baseband unit), an RRU (remote radio unit) and antennas. For 5G, the RRU and antennas are integrated as an AAU (active antenna unit). As shown in figure 4.2.1-1, based on the assumption that 5G will coexist with previous generations, a 5G site will consist of a BBU, an RRU and antennas for 2/3/4G and an AAU for 5G.

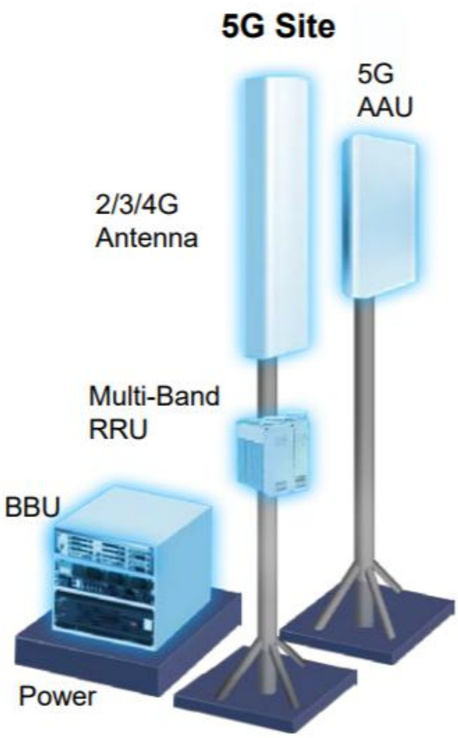


Figure 4.2.1-1: Structure of a 5G BTS[10].

The power consumption of the BBU and the AAU is shown in Table 4.2.1-1. We can see that the energy consumed by the BBU is almost stable, while the energy consumption of the AAU is sensitive to the load.

Table 4.2.1-1: AAU and BBU energy consumption in different loads [https://baijiahao.baidu.com/s?id=1641199237461491515]

Load(%)	AAU average energy consumption(W)	BBU average energy consumption(W)
100	1127.28	293.01
50	892.32	293.01
30	762.43	292.53
20	733.92	293.23
10	699.36	293.41
0	633	293.57

Since the BBU is shared among all generations, it should not be switched off even when the 5G load is low. When we only consider the AAU's energy consumption in different loads, as shown in Figure 4.2.1-2, it appears to be linear with a positive intercept. This means, turning on fewer cells(AAU) with high loads will be more cost efficient than turning on more cells with low loads.

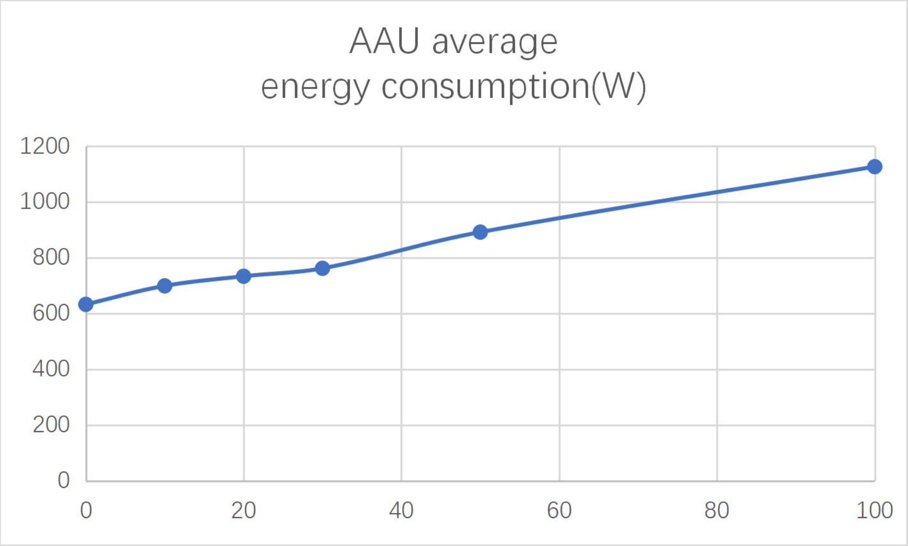


Figure 4.2.1-2: 5G BTS AAU energy consumption in different loads.

As a result, turning on as fewer 5G modules as possible is the goal of the operators in this energy-saving field.

Today, several exabytes (10^{18}) of data are generated per month. The power consumption of base stations is estimated to cost two billion dollars per year. With the growth of the number of sites in 5G era, the energy consumption of these sites is expected to cover 20% of the global energy consumption by 2025[6]. At this volume, any improvement will be significantly beneficial.

In most cases, 75% of the traffic is supported by 30% of the sites [6]. The uses of wireless communication are highly centralized. Meanwhile, the traffic presents a periodic change pattern: high at worktime and low at midnight. This means there is plenty of space to optimize the operation of sites. According to NBD (National Business Daily, nbd.com.cn), Luoyang Branch of China Unicom now set all 5G BTSs to sleep mode from 9pm to 9am every day. In current situation, this action does not do apparent harm to customer experience since the number of 5G users is quite limited, in addition, at midnight, 4G+ can provide similar speed and latency as 5G. Nevertheless, shut down all 5G services is obviously not an ideal long-term policy. Specialists indicate that actions to save energy need to be dynamic and elastic. Saving energy without sacrificing user experience will obviously be the goal. The mesures taken today are kind of too aggressive.

4.2.2 Solution

Nokia is developing a power-saving solution for this problem. The principle is to turn on and off cells/base stations dynamically according to the predicted load.

For confidentiality reasons, concrete non-public data are concealed.

4.2.2.1 Theory

The load of a base station is associated to the throughput generated by end mobile users. In most cases, the loads show a periodic pattern since human activities are regular. Figure 4.2.2.1-1 and Figure 4.2.2.1-2 show the variation of the loads of a base station on Monday and on Sunday. As shown in these two figures, the variation trend of loads is highly similar on Monday(weekday) and on Sunday(weekend). Thus, we can use one single model to match the trend and predict the load in several minutes according to previous load values.

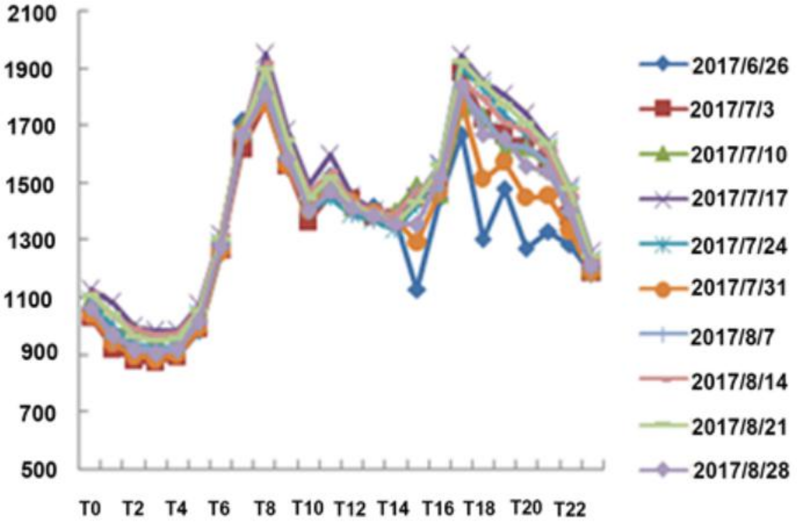


Figure 4.2.2.1-1: Base station loads on Monday. [15]

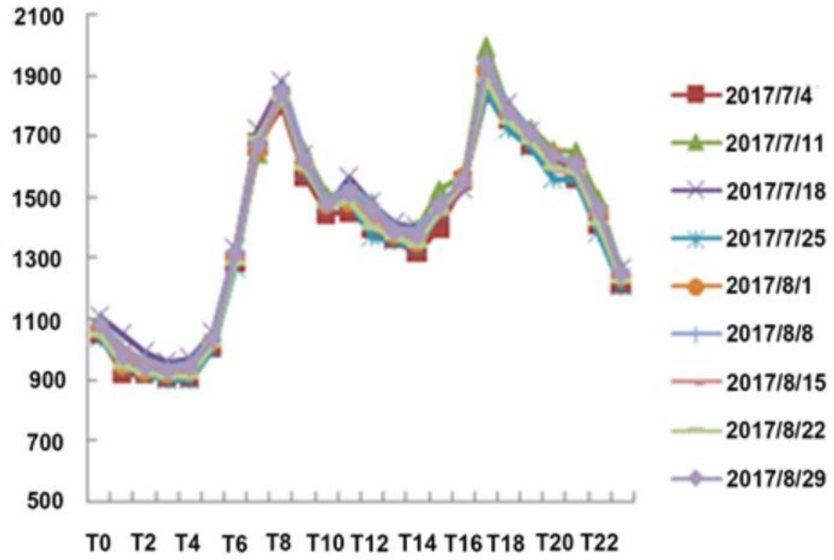


Figure 4.2.2.1-2: Base station loads on Sunday. [15]

Since the data show one-day-period variation trend, we can use machine learning methods to train a model to match the trend then predict future loads based on the model.

After predicting the load of the next time interval, the next step is to compare the predicted value with the threshold. If the predicted value is higher than the upper threshold, a new cell needs to be turned on to assure good service quality; if the predicted value is lower than the lower threshold, a cell can be turned off to save energy.

4.2.2.2 Practical

The main practical part of this solution is to predict the (average) load of the next cycle (time interval) based on the load values of the previous day, as shown in Figure 4.2.2.2-1 below.

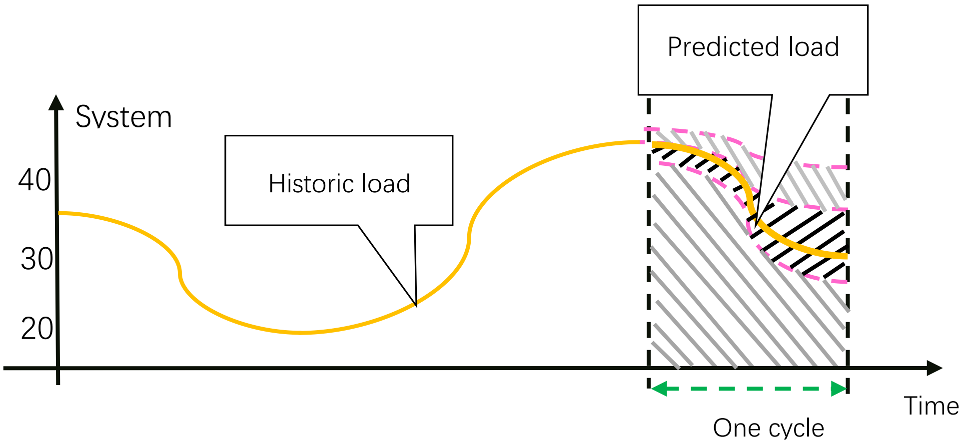


Figure 4.2.2.2-1: Diagram of load prediction.

Deep learning is an appropriate method to solve this kind of problems. Several deep learning methods exist while most of them are based on neural network, including FNN, CNN and RNN etc. In practice, we used a certain deep learning method to train the model. For confidentiality reasons, concrete model structure and parameters cannot be shown in the paper.

The function is initially implemented by using Google's open source deep learning framework: tensorflow. For confidentiality reasons, only a part of the code can be shown.

As we aim at "online training" which means train the model directly on the board, the programming language used is limited. Tensorflow mainly provides python support, but python is not compatible with the board. Thus, we need to implement related functionalities by writing in C and C++.

Although tensorflow provides C APIs, it does not provide related documentation because of which the implementation in C was difficult. After several weeks' research, we eventually found a way to implement online training[1]. The main steps are:

1. Generate the graph(pb) file (and checkpoint files if needed) conventionally by using python

2. Name all necessary inputs and outputs. For example:

```
self.inputs = tf.placeholder("float", [None, self.inputSize], name
= "input")
```

3. In C file, define a model containing all necessary inputs as variables. For example:

```
typedef struct model_t {
    TF_Graph* graph;
    TF_Session* session;
    TF_Status* status;
    TF_Output input, target, output;
    TF_Operation* init_op, * train_op, * loss_op, * save_op, *
restore_op;
    TF_Output checkpoint_file;
} model_t;
```

4. Fill the inputs with data. For example:

```
//set input and output dimensions
const int64_t inputDims[] = { BATCH_SIZE, INPUT_SIZE };
const int64_t targetDims[] = { BATCH_SIZE, 1 };
//set input and output size
size_t inputNbytes = BATCH_SIZE * sizeof(float) * INPUT_SIZE;
size_t targetNbytes = BATCH_SIZE * sizeof(float);
//allocate input and output memory
TF_Tensor* inputs_tensor = TF_AllocateTensor(TF_FLOAT, inputDims, 2,
inputNbytes);
TF_Tensor* targets_tensor = TF_AllocateTensor(TF_FLOAT, targetDims,
2, targetNbytes);
```

```
//fill input and output tensors
memcpy(TF_TensorData(inputs_tensor), train_batch, inputNbytes);
memcpy(TF_TensorData(targets_tensor), target_batch, targetNbytes);
//put input and output into the model
TF_Output inputs[2] = { model->input, model->target };
TF_Tensor* input_values[2] = { inputs_tensor, targets_tensor };
```

5. Start the session and get the result. For example:

```
TF_SessionRun(model->session, NULL, inputs, input_values, 2,
outputs, output_values, 1, train_op, 2, NULL, model->status);
```

The result of C is almost the same as the result of python. This fact proves the feasibility of working with tensorflow in C. More concrete result can be found in subsequent sub-sections.

4.2.2.3 Control flow

When integrating this power-saving feature to the existed base station control system, four components: Use case controller, Data engine, Training engine and Inference engine are mainly involved. The sequence diagram of the two main functions: training (update model) and inference are presented below in Figure 4.2.2.3-1 and Figure 4.2.2.3-2.

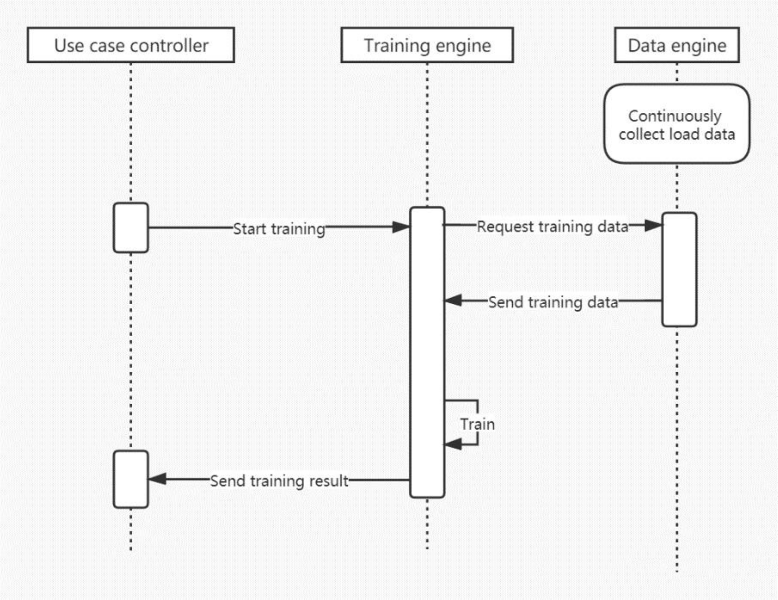
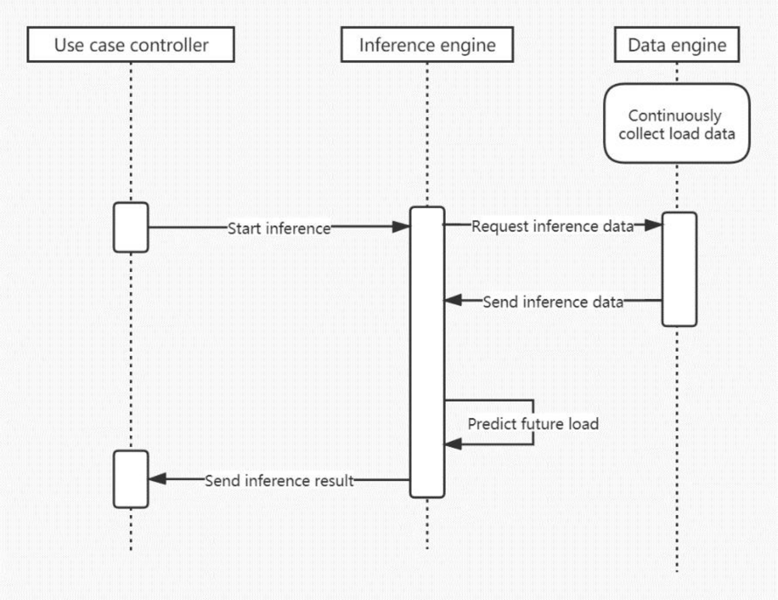


Figure 4.2.2.3-1: Sequence diagram of training process.



4.2.2.3-2: Sequence diagram of inference process.

4.2.2.4 Result

Figure 4.2.2.4-1 below shows the actual loads of a base station for several days and the loads predicted by python on the computer and by C on the board. The figure shows that both C and python can predict the trend correctly and their inference results are highly coincident, although some sudden peaks cannot be forecasted. (Concrete data are concealed for confidentiality reasons)

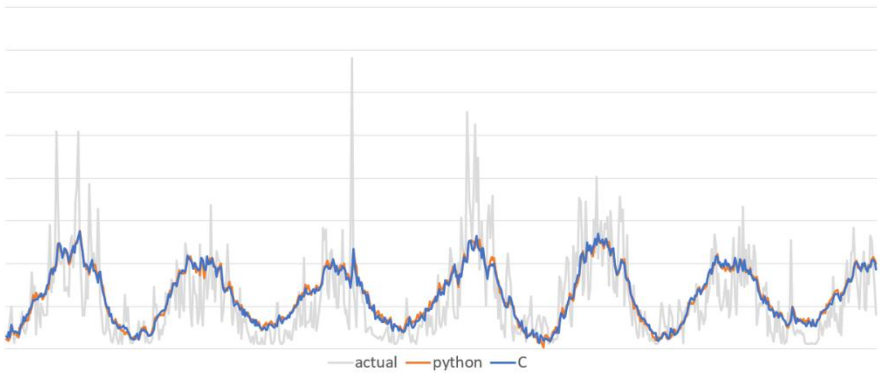


Figure 4.2.2.4-1: Actual load, predicted load by python and by C.

Since the program should not be tested directly on public base stations, we did tests on an internal base station. After integrating this feature with the system already installed on the board, we obtained three-day result below shown in Figure 4.2.2.4-2 (Concrete data concealed for confidentiality reason). To increase test efficiency, the time interval was set as five minutes.

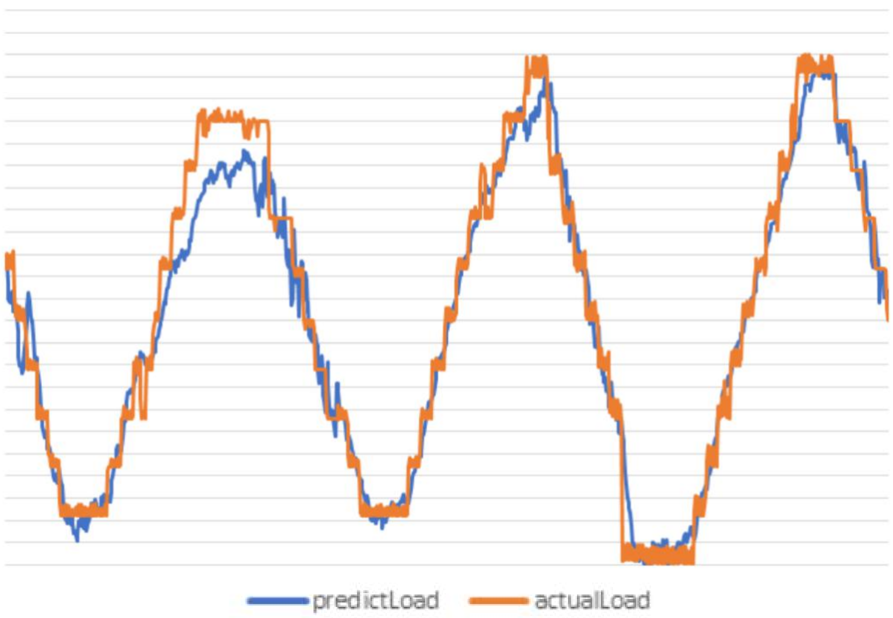


Figure 4.2.2.4-2: Testing results on the board.

When we do not provide any weights information at the beginning, the algorithm shows acceptable result even in the first day and starts to converge in the second day. As the time interval was set as five minutes which was relatively short, the real converge time will be longer, but is still acceptable if the program keeps running.

These simulation results show feasibility that the program be future tested and deployed on real public base stations.

4.2.2.5 Discussion

Our auto-scaling method which is based on the predicted load, compared to the traditional auto-scaling method which is based on the current load, does not save more energy but has better performance theoretically assuming that the prediction is accurate. As shown in Figure 4.2.2.5-1, traditional auto-scaling method always turns on and off cells one interval later than the appropriate time. Thus, to evaluate the gain of this feature, not only the amount of energy saved, but also the variation of the quality of service need to be considered.

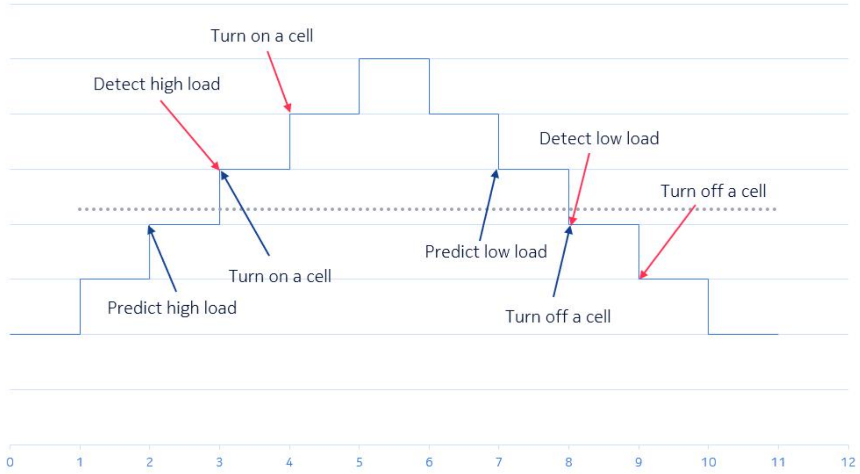


Figure 4.2.2.5-1: Difference between auto-scaling method based on predicted load and current load.

To apply an auto-scaling policy, there must be thresholds. A new cell will be turned on when the predicted load is higher than the upper threshold. Since the energy consumption of the AAU is linear to the load, the upper threshold can be set as 100% in theory if we only consider the power consumption since high load does not imply high power-load ratio. As we cannot ensure 100% accuracy of our prediction, the upper threshold is usually simply set to a certain high value, 70% for example. However, setting such a fixed threshold has a problem. This kind of configuration has an implication that 70% is an “inflection point”: the performance of the base station will drop dramatically when the load exceeds 70% and remains on a good level when the load is below 70%. There is no evidence that such inflection point exists.

Therefore, the decision whether to turn on or off a cell should be made based on the comparison between the energy consumption and the performance of the base station. The gain of the feature comparing to the auto-scaling method based on the current load can be calculated by using the same function in form:

G = f(Esaved, K1,K2,K3…)

Where G is the total gain, Esaved the energy saved and Ki several KPIs.

Possible KPIs:

1. Latency for time-critical use cases

There are many time-critical use cases for 5G like autonomous driving. Their use cases are sensitive to the latency. High base station load may augment the latency for those connections.

2. Packet loss

High load may cause high packet loss rate. This rate can be easily retrieved.

3. User complaints

For daily use cases, the performance loss cannot be measured intuitively since in many cases 3MBps data rate and 2MBps data rate have no apparent difference. Although customer complaints are usually irrational, the number of user complaints can in some degree reflect the general user experience.

Possible calculation methods:

1. Convert all to cost.

Power consumption can be easily converted to electric charge; Increased latency will bring higher risk for accidents which will cause car damage and human injuries, thus increased latency can be converted to equivalent economic loss; Increased user complaints can do harm on the operator's public praise, also necessitates more customer service people, these can be converted to increased operating cost for the operator. The evaluation function can sum all these KPIs to decide whether economically turning on or off a cell is beneficial.

2. Strict requirements.

Sometimes there are strict requirements: the latency must not exceed a certain limit; a certain minimum data rate must always be ensured etc. The configuration of the threshold must satisfy these requirements.

In conclusion, the selection of thresholds should consider both technical factors and social factors. For the upper threshold, the minimum between those calculated by the two methods above should be accepted to ensure the reliability or the high cost efficiency.

By following this approach, the threshold varies for different number of on cells. To be concrete, let the upper threshold be Tmax(i) where i is the number of cells on, then Tmax(1) is based on the comparison between one cell with T load and two cells with $\frac{T}{2}$ load. In the same way, Tmax(2) is based on the comparison between two cells with T and three cells with $\frac{2T}{3}$ load. The threshold will also change when social situation changes, for example the human cost increases or a stricter requirement for latency is set.

To conclude, the selection of the thresholds is tough work. Many factors need to be considered to calculate the best thresholds, meanwhile, these factors especially the social ones like the growth of the number of customer complaints cannot be easily tested or predicted in the laboratory. Thus, much work needs to be done when the feature comes into actual use.

Chapter 5

Conclusion

5.1 Current Challenges of 5G

From ancient times to today, the communication method has evolved a lot, especially recently. In less than fifty years, the telecommunication technology has developed from 1G to 5G, from analog voice communication among humans to digital signal communication among objects.

Great development brings great challenges. We discussed in Chapter 3 that the two most popular specific use cases: autonomous driving and IoT both have strict requirements. Among 5G's three axes: eMBB, URLLC and mMTC, although all of them are important and have many associated use cases, URLLC (ultra-high reliability and low latency) which plays an important role in autonomous driving and IoT industry is the one that necessitates most directly an extreme huge number of cells to be deployed. Only with high density cell coverage can 5G in reality ensure both low latency and high reliability. On this basis, site disguise is also an important challenge in 5G base station deployment since the public's false and irrational fear on radiation can be a great resistance.

Together with the increasing number of base stations, one single 5G base station also consumes more energy than a 4G base station. In 4G era, energy cost is already one of the main parts of operators' OPEX, in 5G era, if no additional efforts were made, the energy cost would increase to an unacceptable level. Thus, we discussed in Chapter 4 the energy-saving solutions for both hardware and software. On the hardware side, more efficient cooling system can be adopted, and the power supply structure can be adjusted; On the software side, dynamic auto-scaling algorithm can be deployed to reduce energy waste when the load is low. With all these measures, the operating costs of base stations can be reduced to an appropriate level for long-term operation.

5.2 Future Challenges of 5G

Although 5G deployment has started and 5G is already in use in some regions, on the macro level, 5G is still in an early stage. There is great uncertainty in the future.

In autonomous driving, 5G is essential to achieve the V2X goal. However, 5G is not the only element that affects the feasibility of autonomous driving, the autonomous driving algorithm and other supporting facilities are important as well. Today's autonomous driving algorithm mainly analyzes the information captured by sensors but implementing V2X is a completely different approach. How and when it will be implemented are now unknown. We expect 5G technology to be mature in the future, expect it to ensure 1ms latency as in theory, however, the reality is usually not perfect. Before 5G and autonomous driving both become mature, 5G is not the only option for autonomous driving. This also applies to other industries including IoT. Before all related technologies including 5G become mature, 5G is not the only option for that industry.

5G will bring high transmission speed and low latency, nevertheless, what industry requires that speed and latency and cannot use wired connection? Although wireless communication is preferable in outdoor environment, there is no situation found now that demands ultra-low latency outdoors except autonomous driving. Similarly, high transmission speed does not seem meaningful for outdoor daily uses. It is indicated that indoor 5G uses will occupy 80% of the total 5G uses, however, indoor use cases can be perfectly satisfied by wired connection and Wi-Fi 6. As shown in Chapter 3.3 Indoor Coverage, for indoor coverage, Ericsson's solution is to cover only hot spots like the conference room by 5G since 5G has low penetrability, Wi-Fi can do the same. In short, now there is no use case that necessitates 5G.

By using a Chinese idiom, 5G is like “chicken rib”: tasteless when eaten but a pity to throw away.

To conclude, 5G is essential for autonomous driving and IoT when they both become mature which will take at least five years. Before that, 5G is replaceable. Thus, in the first several years of 5G, real 5G use cases, like live for 4G, need to be discovered. If no killing app is developed soon, 5G risks of not being able to maintain operation since it has a high operation cost. 5G has a huge advantage compared to Wi-Fi which is that it has a switching protocol while Wi-Fi does not. People need to manually switch from one Wi-Fi to another even when the used one is almost not connectable while 5G does not have this problem. Thus, 5G can try to integrate the whole networking system. When the tariff is lowered to a certain level, maybe for daily uses we will not need wired connection. Instead of connecting to Wi-Fi at home and in the company, we can connect to 5G all the time. This can be an evolution direction for 5G.

Bibliography

1. Asimshankar. (2019). Training TensorFlow models in C. Github Gist. Available from: https://gist.github.com/asimshankar/7c9f8a9b04323e93bb217109da8c7ad2
2. Alriksson, F., Boström, L., Sachs, J., Wang, Y.-P. E., & Zaidi, A. (2020, June 2). Critical IoT for time-critical communications. Ericsson. Available from: https://www.ericsson.com/en/reports-and-papers/ericsson-technology-review/articles/critical-iot-connectivity
3. Bei, W. (2019, March 21). How is the 1ms latency implemented by 5G? Zhihu. Available from: https://www.zhihu.com/question/307958274/answer/712266324
4. Blackman, J. (2021, January 26). 5G and Wi-Fi 6 'blueprint' sets out 'limitless potential' in Industry 4.0, smart cities. Enterprise IoT Insights. Available from: https://enterpriseiotinsights.com/20191016/channels/fundamentals/what-is-mmtc-in-5g-nr-and-how-does-it-impact-nb-iot-and-lte-m
5. Bluespace. (2017, July 12). Achieving a 90% reduction in power consumption for 5G. Available from: https://www.bluespace-5gppp.eu/news/2017/7/12/g4k9pszqqyi1te684kt7b1t1dvzm5m
6. Ericsson. (2020, Mars 11). Why we need a new approach to network energy efficiency. Available from: https://www.ericsson.com/en/blog/2020/3/5g-network-energy-efficiency
7. Ericsson. (n.d.). Bringing 5G networks indoors Whitepaper. Available from: https://www.ericsson.com/en/reports-and-papers/white-papers/bringing-5g-networks-indoors
8. GSMA. (2019, October). 5G, the Internet of Things (IoT) and

Wearable Devices. Available from: https://www.gsma.com/publicpolicy/wp-content/uploads/2017/10/GSMA-5G-IOT-Wearable-Devices-ENGLISH-Oct-2019.pdf

9. Hardesty, L. (2020, April 3). 5G base stations use a lot more energy than 4G base stations: MTN. FierceWireless. Available from: https://www.fiercewireless.com/tech/5g-base-stations-use-a-lot-more-energy-than-4g-base-stations-says-mtn
10. Huawei. (2019, May). 5G Ultra-Lean Site White Paper. Available from: https://www.huawei.com/minisite/5g-ultra-lean-site-2019/pdf_v1.0/5G-Ultra-Lean-Site-White-Paper_en.pdf
11. Huawei. (2020, July 30). NB-IoT Officially Recognized as a 5G Standard, Driving IoT Development. Light Reading. Available from: https://www.lightreading.com/partner-perspectives-(sponsored-content)/nb-iot-officially-recognized-as-5g-standard-driving-iot-development/a/d-id/762804
12. John M, Meredith. (2016, June). 3GPP TR 36.881 V14.0.0. Available from: https://portal.3gpp.org/desktopmodules/Specifications/SpecificationDetails.aspx?specificationId=2901
13. Juda. (2019, July 17). Analysis of how 48V lithium battery be used in base stations. Available from: http://www.juda.cn/news/90407.html
14. Kearney. (2020). 5G: a key requirement for autonomous driving--really? Available from: https://www.kearney.com/communications-media-technology/article/?/a/5g-a-key-requirement-for-autonomous-driving-really-
15. Luo, Y. (2019, April). Urban Base Station Traffic Forecasting Based on Mobile Trajectory. Available from: https://doi.org/10.12677/CSA.2019.94085
16. Mathworks. (2020). FDD and TDD Duplexing - MATLAB &

Simulink. Available from: https://www.mathworks.com/help/lte/ug/fdd-and-tdd-duplexing.html

17. Mudian. (2019, April 24). China's growth in the number of 5G base stations will promote demands in lithium batteries of higher than 155GWh capacity. Sohu. Available from: https://www.sohu.com/a/310068419_100154324
18. Nokia. (2019, May 23). Nokia liquid cooling base station in Satakuntatalo. YouTube. Available from: https://www.youtube.com/watch?v=NIts74-tq_Y
19. Nokia. (2020, June 3). Nokia and Elisa see sustainability leap in world-first 5G liquid cooling deployment. Available from: https://www.nokia.com/about-us/news/releases/2020/06/03/nokia-and-elisa-see-sustainability-leap-in-world-first-5g-liquid-cooling-deployment/
20. Guoming, Pang. (2018, October 8). Kill me if you still don't understand IoT after reading this blog. CNblogs. Available from: https://www.cnblogs.com/pangguoming/p/9755916.html
21. Prasad, R. (2014). 5G - 2020 and Beyond (River Publishers Series in Communications). River Publishers.
22. PSG - Paris Saint-Germain. (2018, November 22). The difference between 4G and 5G on the field with Thiago Silva, Marco Verratti. YouTube. Available from: https://www.youtube.com/watch?v=KoRkxtZJzEI
23. Rembovsky, A., Ashikhmin, A., Kozmin, V., & Smolskiy, S. M. (2012). Radio Monitoring: Problems, Methods and Equipment (Lecture Notes in Electrical Engineering (43)) (2009th ed.). Springer.
24. Scharp, M. P., Scharp, M. P., Persson, O., Persson, O., Scharp, M. P., Scharp, M. P., Persson, O., & Persson, O. (n.d.). Why we need a new approach to 5G network energy efficiency. Ericsson. Available from: https://www.ericsson.com/en/blog/2020/3/5g-network-energy-effic

iency

25. Sohu. (2019, June 29). Nokia's first liquid-cooling base station in the world, highest power in the industry. STM32. Available from: https://www.stmcu.org.cn/article/id-335001
26. Zuojingguantian. (2019, June 6). From 1G to 5G. Sohu. Available from: https://www.sohu.com/a/318911671_150227318911671_150227