Chapter 10
Internet in Your Pocket

Another huge technological development had been happening in parallel to the ongoing digital revolution in mobile phones: the large-scale utilization of the *Internet*.

While the main focus of operators was squarely in optimizing the network performance to match the ever-growing user base for voice calls, a new feature was taking its first steps on digital networks: *mobile data connectivity*.

After all, as the networks already were based on digitized audio data, why not also provide support for generic digital data? The introduction of the *Short Message Service (SMS)* feature had already conditioned users to the possibility of instantly and reliably exchanging text-based messages while on the move, and even some clever applications had been introduced that used the SMS channel as the data connection between the handset and a backend service.

SMS was a kind of late throw-in into the GSM standard, using available bandwidth that was left "free" in the control protocol. By slapping a per-message price tag on a feature that was practically free to implement, the operators over the years have made billions of dollars from this service.

Adding a full-fledged data capability to GSM was kind of a bolted-on hindsight. The early cellular technology had focused on repeating the fixed phone line experience, only enhancing it with the new features that were possible due to the advances in electronics. This was due to the fact that for the incumbent players in the telecoms arena, voice data was seen as being very different from other data traffic.

Fundamentally though, when any kind of data has been turned into a digital form, all data transfers are just streams of bits—voice data differs from other digital data only by its expected real-time performance. Most data-oriented networks at the time did not support any kind of *Quality of Service (QoS)* separation or *traffic prioritization* between different data classes, and therefore the implementation of voice and data traffic in the 2G networks were also handled as two totally separate items.

P. Launiainen, *A Brief History of Everything Wireless*,
https://doi.org/10.1007/978-3-319-78910-1_10

During the early years of the 2G revolution, acoustic modems were the common form of connecting to the Internet. They worked by converting data traffic into audio signals of varying frequencies that were then carried over a normal phone connection and converted back to digital data at the other end of the call. From the system's point of view, you effectively had a normal point-to-point audio call as your data channel, and therefore you also paid by the connection time, not by the amount of data you consumed.

The earliest supported data implementation on GSM was *Circuit Switched Data (CSD)*, which followed the same principle in terms of cost. Similar service had already existed in some of the first-generation networks and copying it was therefore a logical step: the engineers had an existing, working model and simply copied it to the mobile world, repeating the same steps as the second-generation networks became available.

The supported speed of GSM CSD was dismal—9.6 kbps or 14.4 kbps, depending on the frequency band in use, and as mentioned, you still paid by the connection time, even if you did not transmit a single byte over the network. Therefore, CSD was a very poor deal in terms of cost per byte, given the low speeds and the high cost of mobile calls.

Despite limitations, having circuit switched data as a feature did allow general data connectivity on the go, and again, this was a major step forward for some users: being able to check your emails while on the road was a significant improvement, however slow or expensive the procedure ended up being.

By allocating more time slots per user, it was possible to extend the circuit switched speed up to 57.6 kbps through an enhanced version called *High Speed Circuit Switched Data (HSCSD),* which was released as an upgrade. Although this allowed matching speeds with the existing acoustic modems, it was a major problem from the network's point of view, because it statically allocated up to eight time slots for a single user, severely eating up to the total available wireless capacity in a cell. Therefore, for example, in inner cities with a large number of expected simultaneous cellular users, getting full speed for your HSCSD connection was highly improbable: this kind of usage remained so rare that operators preferred to optimize their networks with the ever-increasing number of their voice customers in mind.

To really open up mobile data usage, a better way had to be developed, and *General Packet Radio Service* (*GPRS*), which is often referred to as 2.5G, was the first prominent data extension to GSM. It started emerging on live networks around the year 2000.

GPRS implemented the *packet switched* data model, which was the same as what was the norm in existing fixed data networks. If you activated the GPRS mobile data feature for your mobile account, you no longer needed to allocate a dedicated data channel for the duration of your session, nor did you pay by the minute for the channel usage. Instead of a fixed, costly session where you were constantly pressed on time, you had the feel of being "always on the Internet", and you only paid for the actual amount of data being sent or received.

This was a major conceptual change, as it was now possible to have applications on the handset that were running all the time, using the data channel only when needed and without continuous user intervention.

Prior to the introduction of GPRS, if you wanted to check your email while using a circuit switched data connection, you had to specifically initiate a data call to your mail provider to synchronize your local email buffers. Due to the tedious and often costly process that in many cases just indicated that no new mail had arrived since the last synchronization session, the users only did this only a couple of times per day, potentially missing some time-sensitive mails that actually arrived between the sessions.

With GPRS and its successors, your mail application could repetitively poll the mail server on the background and download and inform you about your new mails as soon as they arrived. This fundamental paradigm shift of continuous connectivity is the core enabling feature of almost all mobile applications today, from *Facebook* to *WhatsApp*, causing the constant, asynchronous interruptions to our lives.

As they say, you gain some, you lose some.

The enhanced speed of GPRS was still based on using several time slots, but it was no longer necessary to allocate these slots for the entire duration of the connection like was the case with HSCSD. The actual, momentary data transmission speed of the connection varied dynamically, depending on the existing load in the base station: the theoretical promise of GPRS was up to 171 kbps, but this was rarely achieved in real-life situations, so compared with the speeds provided by the fixed Internet, even this best-case speed was still horrendously slow, yet it still beat the best circuit-switched HSCSD speeds up to threefold.

The next step up was the introduction of *EDGE*, which was touted as 2.75G, and released in 2003. To show how far the engineers were ready to go in order to make their alphabet soup sound snappy, EDGE stands for *Enhanced Data Rates for Global Evolution*.

The improved bandwidth that EDGE offered was implemented through updating the modulation scheme within the existing time slots. This meant that the higher speeds of EDGE did not eat up more bandwidth but were merely utilizing the existing radio resources more effectively. The maximum promised data speed was 384 kbps—still a far cry from fixed Internet speeds of the time, and as before, the speed depended on the current load on the base station.

Unfortunately, with these second-generation mobile data networks, low and often highly varying data speed was not the only issue that caused poor user experience.

First, let's look at what happens when we browse the Internet:

Common web browsing is based on *Hypertext Markup Language (HTML)* protocol, which is a simple *query-response* scheme: different parts of the web page are loaded through a series of repeated queries to the server. First, the browser loads the core portion of the page, based on the address that the user has chosen for download. It then locates any other references that reside on the page, like pictures or references to external advertisements, continuing to load each of these through additional queries.

This process of going through any remaining references on the page continues until the complete content of the originally requested page has been downloaded, and hence the overall complexity of the page defines the total amount of data that needs to be loaded. The overall time needed for this is then directly proportional to the available data transmission speed.

On top of this, each additional query-response also has to go from your device to a server somewhere in the world and back, and your actual physical distance to these servers will add delay to each roundtrip. In optimal cases, when you are near to the data source in a well-connected, usually urban setting, you may spend only 0.02 seconds extra time for each roundtrip, meaning that even a complex web page will load all its parts in a second or two. But if you are in the middle of nowhere, with poor connectivity, residing far away from the server you are connecting to, this extra time spent on each required query-response can go up to 0.5 seconds or even more.

To add to the complexity, the definition of "far" depends wholly on the *topology* of the network infrastructure you are connected on, not your actual physical distance. As an example, I'm writing this particular chapter in Manaus, which is roughly 8,000 kilometers away from London as the crow flies, but the only data connection from Manaus goes to the more industrialized part of Brazil, down south, and from Brazil, data packets going to Europe are routed first to the United States.

Therefore, according to the network trace I just performed, if I access a server in London, my data packets go first 2,800 kilometers south to Rio de Janeiro, continue 7,700 kilometers to New York, and from there another 5,500 kilometers to London. Therefore, my apparent 8,000-kilometers direct distance has now been doubled to 16,000 kilometers, and one such roundtrip consumes about 0.3–0.5 seconds of time, depending on the dynamically changing network load along the way.

This delay is solely based on the infrastructure you are connected to and is unavoidable. Part of it is due to the physical limit of the speed of light, which also limits the speed of electrons or photons making their way across the data networks, but most of it is the overhead needed to transfer these data packets across all the routers between my location and the source of the web page I am loading.

In my example, about 0.11 seconds of my roundtrip time is due to the laws of physics. That is the delay caused by the fact that my data packets need to traverse the physical network distance of 16,000 kilometers twice during each query-response roundtrip.

But more importantly, on their way between Manaus and London, my packets jump across 30 different routers, and they all need some minuscule slice of time to determine where next to send the packets.

The overall delay spent on routing and transferring the data is called *latency*, and because I was doing my example above while connected to a wired network, this was the best-case scenario. When using mobile data networks, you are facing another layer of latency, as the implementation of the air interface adds its own delay for receiving and transmitting data.

In second-generation networks, this could easily add up to one second to the processing of each single query, far exceeding the above-mentioned, unavoidable delays. Therefore, if the web page that is being loaded contains lots of subqueries,

the perceived performance difference between fixed and 2G mobile networks becomes very noticeable. Hence, the further development of mobile data networks in terms of providing a better data experience was not only about data connection speed, but just as much about reducing the system-induced latency.

In terms of the available data speed, EDGE provided much more than would be needed to transfer digitized audio as pure data across the mobile network. Latency and lack of Quality of Service were still issues, but with EDGE, it was now possible to have decent quality audio connectivity purely in the data domain. The quality that was achieved with these kinds of *Voice over IP (VoIP)* services was still not on the same level as a dedicated digital voice channel, but thanks to the advances in data connectivity and achievable data speeds, the separation of audio as a special case of digital data over other types of data started to look like an unnecessary step.

Although this was not yet an issue with the 2G networks, operators had started to worry about improvements that were planned for the next generation systems: their concern was about becoming a pure *bit pipe*—if mobile data progresses to a point in which it has low latency and predictable Quality of Service, it would become possible for any greenfield company to set up VoIP services that use the operators' data networks only as a data carrier. In the worst case, operators, who had invested hundreds of millions in their network infrastructure, would wind up just transferring the data packets for these newcomers, and the customers would only look for the lowest data connectivity price, causing a race to the bottom and crashing operator revenues.

Currently, with the high-speed, high-quality fourth-generation *(4G)* data networks and eager new players like *WhatsApp*, this threat is becoming a reality, and it is interesting to see what will be the operators' response to this. Some operators have accepted this as an unavoidable next step, and are actively working on becoming the "best quality, lowest cost bit pipe" for their customers, while others, with considerable voice revenue, are looking for ways to fight this transformation.

In the current landscape of fixed price monthly packages that contain static limits for mobile call minutes, mobile data amount and the number of text messages, it is interesting to note that some operators have already included particular data traffic like *Facebook* and *WhatsApp* as unlimited add-ons. As *WhatsApp* now supports both voice and video calls, and often appears to provide much better voice quality than the up-to-the-hilt compressed mobile voice channels, giving a free pass to a competitor like this seems rather counterintuitive.

But during the early days of second-generation mobile data improvements, this potentially groundbreaking paradigm shift to all-encompassing data traffic was still a future issue: the operators had other things to worry about.

Apart from speed and latency, the screens of the handsets were a far cry from the displays that users had become used to in the fixed Internet, and to add to the insult, the available processing power and memory capacity of the handsets were tiny compared with personal computers.

But where there's demand, there will be supply.

The operators realized that enabling easy-to-use mobile data traffic would open yet another, very lucrative revenue stream, and the existing limitations of portable

hardware had to be worked around to lure the customers in. Instead of supporting the existing, complex HTML protocol that was and still is the root of all Internet browsing, a much more limited set of page management commands had to be taken into use.

The first introduction of such scaled-down mobile Internet happened in Japan with *NTT DoCoMo's i-mode* service that was introduced in 1999. Despite its early mover advantage and rapid expansion to cover seventeen countries, i-mode eventually ended up failing outside of Japan. This was partially caused by *Nokia*, which did not warm to this standard that directly competed with their ongoing mobile data activities around *Wireless Application Protocol (WAP)*.

WAP was *Nokia's* first step towards the promise "to put the Internet in every pocket" made by the CEO of *Nokia*, Jorma Ollila, on the cover of the 1999 *Wired* magazine. This happened at the time when the media buzz around these kinds of grandiose promises was high, sometimes popping up in the most unexpected places: in an episode of *Friends*, Phoebe sits on a couch at *Central Perk* café, reading that particular *Wired* magazine with the cover picture of Mr. Ollila clearly in sight.

As was indicated by this kind of penetration into the Pop Culture, *Nokia* had already become the undisputed 800-pound gorilla of the mobile marketplace, and didn't hesitate to use its strong position to dictate the direction of the industry. *Nokia*'s R&D was already struggling to cope with the multitude of existing differences in their mainstream markets, so adding i-mode to their product line mix would just muddy the waters even more. Only the minimum effort was therefore dedicated to handsets supporting i-mode.

As a result, any early adopters of i-mode were forced to use handsets that were coming only from Asian manufacturers, and the colorful, toy-like approach that worked in Japan and a handful of other Asian markets did not migrate too well to the rest of the world. Another blow came when the two other major operators in Japan chose WAP, and so there was no strong reason for *Nokia* to aggressively extend their activities to i-mode. *Nokia*'s brief entry to i-mode compatible devices only happened six years later, with the introduction of *Nokia N70 i-mode edition* for the Singapore market. At that time, the emerging mobile data revolution was already rapidly making both i-mode and WAP obsolete.

In the meantime, WAP went through a series of incremental improvements. Small extensions, like *WAP Push*, that allowed the handset to be notified of asynchronous events like the arrival of email without actively doing a repetitive polling, truly and positively improved the perceived user experience. To pursue the possibilities for both handset sales and additional network revenues, both the device manufacturers and the operators created a massive marketing campaign around WAP.

Unfortunately, the actual services hardly ever lived up to the expectations: typically, the operator acted as a gatekeeper to all WAP content, actively trying to isolate customers from the "true" Internet. Operators wanted to extract maximum value from the content that was provided and keep it fully in their control. For the developers of content, this created an extra hurdle as they had to gain access to these *walled gardens* that had been deliberately set up by the operators.

The driver of this walled garden idea was the success that *NTT DoCoMo* had had with their i-mode standard in Japan. The value-added services in i-mode were strictly controlled by *NTT DoCoMo*, and the cash flow that *NTT DoCoMo* was raking in proved to other operators that a similar approach would be the way to go with WAP, too. Operators held the keys to their WAP gardens, which meant that every deal by content providers had to be negotiated separately instead of just "putting it on the net" for the customers to access, like is the *modus operandi* for the fixed Internet. As a result, the available WAP services tended to be very fragmented—if you changed to another operator, the set of value added services that were provided by the competing operator was very different.

Naturally some operators saw this as an additional benefit, as it sometimes stopped otherwise unhappy customers from changing their network operator.

Despite all frustrations caused by slow speeds and small screens, customers took their first baby steps with the mobile Internet, and the introduction of these often very simple services did again create a totally new revenue stream for the operators. *Ericsson*, one of the early providers of both handsets and networks, very descriptively called WAP "the catalyst of the mobile Internet", and in hindsight, both i-mode and WAP were necessary and unavoidable steps. They simply aimed at squeezing the maximum value out of the very restrictive technical limitations of the time, preparing the users for the incoming mobile data revolution that could be predicted by the expected advancements in technology in the not-so-distant future.

The WAP era also did make a handful of early application development companies momentarily very rich, although this was seldom due to the actual revenues from their products. The most prominent cases were due to the enormous hype around the "WAP boom", which was part of the wider "dot-com boom" that was engulfing almost anything that had something to do with the Internet. The hype pushed the value of the stocks of WAP-related companies sky high before crashing them down again—just like what happened with the "radio boom" in the early years of the 20th century, except now with some extravagant, often very public indulgence during the short lifespan of these companies.

Good examples of companies in this category were the Finnish WAP-application developer *Wapit*, and another Finnish mobile entertainment company *Riot-E*, both of which offered a brief but very expensive introduction to risky investing around the turn of the 21st century. *Riot-E* burned through roughly 25 million dollars in its two years of existence, and not a cent of that came from the founders of the company.

Adding WAP-capabilities to handsets was a boon for manufacturers and operators, as it accelerated the handset renewal cycle and hence kept the sales going. Despite the limited set of available WAP features, just being able to seamlessly access their email on the go was good enough reason to invest in a brand new WAP phone for many mobile handset users.

Along with these early baby steps to mobile data, the basic voice usage also kept on growing steeply, and the existing second-generation networks started to saturate under the ever-growing load.

History was repeating itself.

There was no way to cram more capacity from existing frequency bands, and the improvements in both data processing capacity and display capabilities increased

the need for mobile data services. As various wireless data services started growing exponentially, mobile voice connectivity was no longer the only game in town.

The only way to move past the current limitations was another generation change, so it was time to take a look at the new developments in the manufacturers' laboratories, combine them with all the learnings gained from the field, and start pushing for a *third-generation (3G)* solution.

Although the scientists and engineers involved in the standardization process were hard-core professionals with very objective viewpoints, striving to find the best possible solution to the task at hand, there were enormous, underlying corporate interests connected to this work: whoever could get their designs incorporated into the new standard would gain huge financial benefit from the future licensing of their patents.

The standardization work tried to balance this by seeking ways to incorporate the best ideas into the new standard in novel and improved ways that would bring in the benefits with minimum cost impact in terms of existing patents, but in practice those companies with the most effective and well-funded research departments would also be able to drive this work, patent the new proposals made by their research teams, and hence ensure their share of the future licensing income.

One outcome of the US "free market" experiment was that the CDMA technology had grown to have a considerable and established user base in the United States. Despite the huge success of GSM, some theoretical improvements that could be gained by using the CDMA technology had now been proven in practice, and CDMA had its own upgrade path, called *CDMA2000 Evolution-Data Optimized (EV-DO)*, which was being pushed actively as the next generation CDMA for those operators who had CDMA in their networks.

At the same time, the holders of GSM patents were painfully aware that GSM could not offer a step up to 3G without considerable modifications: GSM, having been being early in the game, inevitably had some rough corners that were not optimized to the maximum possible level.

What was needed was a joint effort to find a solution that would avoid the kind of fragmentation that happened in the USA with the 2G networks. As a result, the maker of CDMA, *Qualcomm*, and the largest patent holders of GSM, including *Nokia* and *Ericsson*, together with other active companies in the wireless domain, created teams to work on the next generation standard.

Entire floors were booked in hotels for the herds of engineers and lawyers of the competing wireless technology companies for the duration of these negotiations, and when the dust finally settled, a proposal that combined the best parts of the competing technology directions had been forged.

The new official name of this amalgamation was *Wideband CDMA (WCDMA)*, and the major patent holders turned out to be *Qualcomm* with the largest share, followed by *Nokia*, *Ericsson*, and *Motorola*. It preserved the flexibility of SIM cards and the requirement of inter-operable, non-proprietary network elements of the GSM system, while adding some fundamental technological improvements offered by the CDMA technology.

Hence, despite its failure to gain major traction on a global scale, CDMA succeeded in incorporating their *air interface* to the global 3G standard. This was a major win for *Qualcomm*, helping it to grow into the prominent communications and computing hardware company it is today: at the turn of the 21st century, *Qualcomm*'s revenues were about 3 billion dollars, growing fifteen years later to over 25 billion—almost 20% annual compounded growth, and the largest money maker for *Qualcomm* in recent years has been its technology licensing business.

Just before this book went to print, Singapore-based *Broadcom Ltd*. made an unsolicited offer to buy *Qualcomm*, with a price tag of 130 billion dollars, including *Qualcomm's* existing debt. Although this was the largest ever tech takeover offer, it was rejected by the board of *Qualcomm* as being "way too low". Subsequent, even higher offer was eventually blocked by the U.S. government on national security grounds.

It just goes to show how the wireless industry has created huge value out of the effective harnessing of the electromagnetic spectrum just in the past two decades.

Similarly, although *Nokia* no longer makes handsets, according to an agreement made in 2011 and renewed in 2017, it still receives a fixed sum from every *iPhone* sold, thanks to the essential patents it holds for the cellular technology. Although the actual details are not published, with hundreds of millions of devices being sold annually, this is a major revenue stream: in 2015, the division of *Nokia* that receives related patent income from *Apple* and other licensees, indicated revenues of about billion dollars, and when the agreement was renewed in 2017, a two-billion-dollar additional one-off payment was reported.

Hence it is no wonder that new patent wars pop up all the time:

Sometimes these disputes seem to be on very different levels, though: *Apple*, a latecomer to the mobile phone business but currently making the biggest profits off it, doesn't have any significant essential patents to push. Therefore, it is not covered by the *fair, reasonable and non-discriminatory (FRAND)* agreements, where the incumbent patent owners volunteer to cross-license their standard-essential patents with lowered cost amongst the group that worked on the standard.

Instead of these patents that are essential to the actual operation of their handsets, *Apple* successfully milks their design patents of "rounded corners" and other less technical things, like the "slide-to-unlock" patent—a virtual imitation of a thousands of years old physical latch approach. Copying that basic concept ended up costing 120 million dollars for *Samsung* in 2017.

Despite its apparent simplicity compared with essential patents that many other wireless pioneers hold, this approach seems to work well: at the time of writing this, *Apple* announced that they have 250 billion dollars in the bank, and have just spent 5 billion in their new "flying saucer" headquarters in Cupertino—the most expensive office building ever.

As with every generational change in network functionality, the network hardware manufacturers try to find new, exciting use cases for the new functionality that is expected to be available in the incoming upgrade. One such promise that was much touted in relation to 3G was the support of *video calls*. This feature was prominent in all early marketing material describing what nice things were to be expected, but actual market studies soon indicated that it really was not a feature that the customers were waiting for.

Fast mobile data? Yes, please. Video calls? Not really.

Another generation switch would again be astronomically expensive, so the network manufacturers were somewhat worried about their future prospects, yet in the end, 3G turned out to be an easy sell, because for the operators the biggest issue with 2G was the severely saturated networks in densely populated areas.

So out went the futuristic Dick Tracy-video calls—the major selling point in the end for 3G was the good old voice service. It was all about matching network capacity with the growing customer demand, and despite all other advantages of 3G, enhanced mobile voice capacity turned out to be the key feature that kept the renewal cycle going. 3G also improved the encryption scheme used for the connection: the original 2G encryption method had been made practically obsolete by advances in computing power, and it was also possible to configure the 2G base stations to transparently disable the encryption.

Although there was now a clear path to interoperable 3G globally, China still threw a spanner in the works. As a single market of almost 1.5 billion customers, China is large enough to dictate their own policies as they please, and in order to support their own technology research and reduce the amount of money needed to be paid to the WCDMA patent owners, China mandated that the state-owned national operator, *China Mobile*, would be running a homegrown 3G standard. This was created with the not-so-catchy name of *Time Division-Synchronous Code Division Multiple Access (TD-SCDMA)*, and during the roll-out, it still used the standard GSM as the secondary layer for backwards compatibility.

Today, *China Mobile* is the world's largest operator by far with 850 million subscribers at the time of writing this. Hence although the TD-SCDMA standard never made it outside of China, it has a large enough user base to support very active handset and base station development.

Faster data was one of the fulfilled promises of 3G, and it has gone through a series of improvements in terms of data transmission speeds after its introduction, yet as the implementation and tuning of various features has a lot of flexibility, there is plenty of variation regarding what kind of speeds can be achieved in different operator environments.

Another important goal was to reduce the extra latency of the wireless air interface, and 3G got it down to the 0.1–0.5 second range.

And development did not stop at 3G.

The switch to *fourth-generation (4G)* systems has been progressing rapidly in the last couple of years, through a process called *Long-Term Evolution (LTE)*. This has the goal of increasing data speeds again to about five times faster in real-life situations as compared with basic 3G, and now the new generation roll-out also addressed the *upload* speed, which means that sending data like digital pictures from your smartphone is now considerably faster than with 3G.

As the LTE switch is a more gradual upgrade than what happened during the change from 2G to 3G, many operators call it *4G LTE*, which more realistically describes what is actually on offer.

China Mobile has also rolled out their extension to the TD-SCDMA standard, called *TD-LTE*.

For a while, another standard, called *Worldwide Interoperability for Microwave Access (WiMAX)* was pushed as a potential contender for LTE. It has its roots in the work done for the South Korean *Wireless Broadband (WiBro)* network that was launched in 2006.

To push WiMAX, *Samsung* joined forces with *Intel Corporation*, which had otherwise utterly missed the mobile revolution in its core processor business:

The new king of the "mobile hardware hill" was a British company called *ARM Holdings* in Cambridge, and as a major deviation of the norm of computer chip manufacturing, *ARM* does not manufacture its low-power/high-performance processor designs. Instead it licenses them to other manufacturers, like *Apple* and *Qualcomm*, and even to *Intel*, which finally licensed the ARM architecture in 2016. Not having to invest in expensive semiconductor fabrication plants turned out to be very profitable for *ARM*: 75% of *Arm Holdings* was bought out by the Japanese *Softbank* in 2016, with a price tag of over 30 billion dollars, meaning that the average value of this 26-year-old company grew by more than 1.5 billion dollars per year.

Despite having major players like *Samsung* and *Intel* behind WiMAX, the economies of scale were not in favor of another high-speed wireless standard that had no obvious prior user base. *Sprint Nextel* in the United States was the most prominent operator that was originally planning to have dual-mode EV-DO and WiMAX as their network infrastructure, but they eventually scrapped their plans and selected LTE instead.

WiMAX as a technology is still alive and well, however: it exists in special installations around the world, aimed at replacing fixed Internet connections.

With the increasing speeds and capacities, LTE has now progressed to a point at which it is seriously contesting fixed Internet connections. Rolling out this kind of *wireline replacement* offers a strong business case for operators: maintaining cables for fixed Internet is very costly, especially in rural setting with long cable runs and lots of thunderstorm activity. Hence, replacing a myriad of copper cables with wireless 4G LTE modems reduces projected long-term maintenance costs considerably.

Naturally, for places like emerging markets that have zero existing fixed Internet infrastructure, LTE offers a way to quickly provide fast data access to millions of potential users, meaning that a country's telecoms infrastructure can go from zero to state-of-the-art in less than a year. The potential for improved economic growth after such a leap forward is huge.

Unless your usage profile for some reason requires speeds of 100 Mbps or more, 4G LTE should be more than adequate for even those urban dwellers addicted to 24/7 video streaming, and the shift from wired to wireless has been accelerated globally after the widespread roll-out of LTE networks.

As an example, I upgraded my parents' rural fixed Internet connection to LTE-based connection, going from 500 kbps uplink + 2 Mbps downlink to 5 Mbps uplink + 30 Mbps downlink, for half the cost of the earlier fixed connection. This was one of the rare deals in which everyone was happy: the operator got rid of an error-prone, several kilometers long copper connection that had already been fried

twice by a thunderstorm, and the customer saw considerable improvement in service quality, whilst still saving on the cost of the service.

One significant improvement that has vastly reduced the latency in LTE networks is the *System Architecture Evolution (SAE)* that streamlined the backend side of the networks. As a result, SAE flattened the structure of various network elements that were present in the 3G standard, reducing the overhead needed to process the transmitted and received data packets. Thanks to a low latency of under 0.1 seconds, 4G LTE is good enough even for online gaming.

And we haven't seen the end of this road yet.

As the electronic circuitry improves and processors get faster, the drive is to move to higher and higher frequency bands, allowing more bandwidth for the data.

The *fifth-generation (5G)* systems, that are just in the first deployment phase, aim to further reduce the latency and provide a ten-fold increase over the current 4G LTE network speeds. This moves the available data transmission speeds to the *gigabits per second (Gbps)* range, matching or surpassing the common fixed connection speeds at home.

As discussed in TechTalk ***There's No Free Lunch***, the higher the frequency, the more room there is to modulate, and one reason for the new, higher data speeds with 5G is exactly due to the use of much higher frequency bands, which initially will cover 24–28 GHz frequencies. And for the first time, 5G is expected to allow the magic trick of supporting *full-duplex* traffic on a single frequency, therefore potentially doubling the available network capacity.

Another emerging technology that is already present in some LTE implementations is the intelligent, *multiple-input, multiple-output (MIMO)* antenna solution. This will enable dynamic *beamforming*, which means electronically changing the radiation pattern of the antenna to match the observed direction of the counterparty of a wireless connection, hence focusing the transmitted signal to the active direction only. If this trick is performed both at the base station and the handset ends, it creates a high-energy "signal bubble" at both locations, reducing interference, extending the perceived coverage area and potentially allowing a much tighter frequency reuse scheme to be used. And when the same frequencies can be reused more effectively, a single base station can handle more simultaneous connections, removing the need to install additional base stations for heavily congested areas.

This will be the standard approach for 5G: early in 2018 *Nokia* announced their new *ReefShark* 5G chipsets, which are expected to provide a throughput of up to 6 terabits (Tbps) for a single base station. To put this in perspective, this is 100 times the recorded overall data traffic at the stadium during the 2017 *Super Bowl* Football Championship game in Houston.

The MIMO technology supported by these new chipsets enables the reduction of the size of antenna structures, while at the same time they cut down the antenna setup's power consumption by two thirds as compared with conventional solutions.

Similarly, the drive in 5G is to cut down the overall power consumption of the wireless connection, therefore making it possible to connect low-power, intelligent devices directly to the existing 5G networks.

All in all, the step to 5G will be another true generational change, and it heralds the ongoing major shift from fixed Internet use to the wireless Internet: in 2017, the

end users already spent 70% of their Internet-related activities on their wireless devices—an earthshaking shift that was initiated by the humble i-mode only 18 years ago.

What used to be called "mobile Internet" is becoming just "the Internet".

The world has come a long way from the early spark-gap technology, but if Marconi, Hertz or Tesla could have been transported into our time, none of them would have had any problem understanding the principles behind our latest and greatest cellular networks. Although the technology has leapfrogged, the fundamental concepts remain the same.

And all this development is still happening through conventional means: new protocols are defined, they are tested with extremely expensive, often bulky, purpose-built initial hardware, and when issues have been fixed, new purpose-built microchips are designed, paving the way to consumer-priced, portable products.

Integrating multiple functions on a single microchip allows lowering the cost and power consumption, thus pushing the technology to a price point that is needed for true mass market products.

In addition to this, when multiple wireless identities, like GPS, Wi-Fi and Bluetooth are combined into a single microchip, various interference and timing issues can be avoided by optimizing the shared logic behind these parallel operations.

Wi-Fi and Bluetooth are discussed in Chapter 11: ***Home Sweet Home***.

For an emerging, vastly improved way of speeding up new developments in the utilization of the radio spectrum, see the description of *Software Defined Radio (SDR)* in TechTalk ***The Holy Grail***.

Thanks to all this continuing progress in wireless technology, we consumers can seamlessly enjoy our *cat video of the day* on our smartphones, balancing the anxiety created by all of those *Instagram* shots sent by our friends and coworkers on their seemingly fabulous and entirely trouble-free holidays.

The final limiting factor is the availability of frequencies for all this use, and as discussed earlier, these frequency bands are a limited resource, strongly regulated by each country. When 3G was defined, it reused parts of the existing spectrum that was already allocated for mobile phone use, but also expanded to new, higher frequencies that were not in use yet. Part of this was due to the fact that making hardware that worked on higher frequencies used to be very costly, but had become available due to the advances in electronics.

The way these new frequencies were taken to use varied widely across the world: some countries simply allocated them for the local operators, but other countries saw the opportunity to milk this new "access to the airwaves" and opened special *spectrum auctions* for the operators to compete in.

And in many cases, these auctions turned into an incredible frenzy. For example, the government of Germany gained over 50 billion dollars off their radio spectrum auctions, whereas the United Kingdom added roughly 34 billion dollars to the government's coffers. This auction caused a serious hangover amongst the operators in Great Britain, as all this money was a sunk-in cost, separate from the funds desperately needed to purchase and deploy the required 3G infrastructure.

Unfortunately, the effects of this kind of "government leeching" can have very long-lasting negative consequences: according to a study done by Britain's *National Infrastructure Commission* in late 2016, the newest 4G LTE network in the UK was on the same level as in Albania and Peru, and at least both of them have complicated, mountainous terrain to deal with.

Hence, while the dwindling financial situations of governments can be nicely topped up by such "air sales", the citizens of these countries end up having slower deployment of the new services.

Government-mandated rules are not always bad, though: a major catalyst for intra-operator competition was the legal requirement that forced the operators to allow customers to keep their existing mobile number while changing to a competing operator. This approach has now been deployed in all major markets, and the induced potential for *churn rate*, which is the number of customers that switch to a competitor, keeps the operators on their toes, for the benefit of customers.

With available frequencies a limited resource, old services are decommissioned in order to release existing frequency bands. In the ongoing transition from analog to digital television, as was discussed in Chapter 5: ***Mesmerized by the Moving Image***, the main driver is to release frequencies formerly used for television broadcasts to new uses.

As many parts of the television spectrum conveniently reside right next to existing mobile phone bands, the expectation is that these can be used to further increase the available mobile communications capacity. The auctions that happened as part of the 3G deployment have clearly shown the potential value that this might bring to the cash-strapped governments of the world.

A recent example of an auction of reallocated television frequencies happened in the United States in early 2017. Although one of the winners was the established mobile operator *T-Mobile*, the two others were *Comcast* and *Dish*, which were approaching the wireless space from a very different direction, being cable television, satellite television and Internet connectivity providers. At the time of writing this, *T-Mobile* have already announced that they will use the acquired frequencies to improve connectivity in rural areas.

This recent auction shows that the feeding frenzy seems to go on unabated—the total sum paid for these new frequency bands in the United States was almost 20 billion dollars.

Although the new frequencies will eventually expand the available wireless services, it is us customers who will finally foot the enormous auction bills, in one form or another.

Apart from the auctions, there have also been other activities aiming at benefiting from the fact that each and every television channel is not in use at every possible location and at any given time. The concept of utilizing allocated but unused television channels is called *White Spaces*, and in this context, the non-active channels can be used for unlicensed wireless Internet access, provided that the users of this technology consult a dynamically changing *geolocation database*.

This database dictates the possible frequencies that can be repurposed in a certain area at a certain time. Such databases were first defined for the United States, the United Kingdom and Canada, and those countries have also seen the first White Space-based installations go online.

The technology utilizing the White Spaces has been snappily named *Super Wi-Fi*, although it has very little to do with the actual Wi-Fi standards that are discussed in Chapter 11: ***Home Sweet Home***. Speed-wise, Super Wi-Fi is not in any ways "super" to Wi-Fi, and with its currently implemented top speed of 26 Mbps, it is also slower than 4G LTE and WiMAX, but for residents in a sparsely populated area where there are no other means to access the Internet, it provides a reasonable alternative for getting online. By using lower frequencies than the other standards, the coverage area of an individual Super Wi-Fi base station is much larger, so the deployment costs are lower than for the competing standards.

The White Spaces-concept is a great example of how a combination of modern data processing, users' geolocation information and wireless technology can be utilized to create new services that would otherwise be impossible due to existing, inflexible frequency allocations. Dynamically reusing television channels that are not in use at any given time makes it possible to maximize the utilization of our limited wireless resources, ultimately benefiting both the end users and the service providers that are happy to jump to support a niche market like this.

Our constant quest for improved bandwidth keeps on moving towards higher frequencies, while at the same time repurposing frequencies that have become obsolete due to advances in digital technology, or being "multiplexed" in terms of time and space, thanks to advances in computer technology. And as has been the case with the speed and capacity of computers, whatever improvements and advances are made, they will be utilized in full a couple of years after their initial deployment.