**Chapter 3 Link Budgets and Latency**

**3.1 Latency and 5G Standards**

Part of the purpose of writing this book is to provide engineers and product planners in the satellite industry with technical detail on the 5G standards process including the New Radio work on the physical layer and the related relevance to next generation satellite service development. Conversely we want to provide the 5G community with visibility to the performance potential of satellites for a number of 5G use cases including counter intuitively some latency critical use cases.

In this chapter we look specifically at the parts of the 5G standards process that have an impact on latency across the four presently designated 5G application domains, enhanced mobile broadband (eMBB) , Low Mobility Large Cell (LMLC), Ultra Reliable Low Latency Communications (URLLC) and Massive Machine Type Communications (MMTC)

You might think that the latency story would only be relevant for URLLC but in practice URLLC services can be delivered as a pre-emptive payload within an eMBB channel, low mobility large cells need to take into account flight time delay from the base station to the user and IOT device and back again and the users and or devices could be moving at 1000 km/h (aircraft) or 500 km/h (trains).

The most extreme latency requirement in the 5G use cases (which we cover in more detail in Chapters 9 and 12) is for MMTC IOT connectivity in ‘the factory of the future’ with a minimum value of 100 microseconds.

In a 5G network it is also important to differentiate between user plane latency and control plane latency.

**User plane latency** is the contribution of the radio network to the time from when the source sends a packet to when the destination receives it (in ms). It is defined as the one-way time it takes to deliver an application layer packet/message from the radio protocol layer 2/3 service data unit (SDU) ingress point to the radio protocol layer 2/3 SDU egress point of the radio interface in either uplink or downlink in the network for a given service in unloaded conditions, assuming the mobile station is in the active state. The minimum requirements for user plane latency are 4 milliseconds for eMBB and one millisecond for URLLC assuming unloaded conditions (i.e., a single user) for small IP packets (e.g., 0 byte payload + IP header), for both downlink and uplink.

**Control plane latency**  refers to the transition time from a “battery efficient” state, for example an idle or deep sleep state to the start of continuous data transfer – in effect the time between being asleep and active. The minimum requirement for control plane latency is 20 milliseconds though there are arguments for reducing this to 10 milliseconds or less. The purpose is to reduce power drain in battery driven devices and energy cost and energy consumption in the network. In IOT applications there can also be a defined time period between wake up events for example if a life of ten years is required from a button cell battery, a device might only wake up every few hours at a defined moment.

Control plane latency also determines the length of time it takes the network to respond to changes in loading condition. So for example if offered traffic is bursty, the traffic offered both at the radio layer and network can vary dramatically and rapidly. In an ideal world, radio and network bandwidth would be provisioned to accommodate the most extreme loading conditions but this would mean the radio layer and network would be underutilised for most of the time. In practice, traffic is buffered and this introduces delay and delay variability. If there is insufficient buffer bandwidth at any point then packets will be lost or discarded and will need to be retransmitted. In effect we are saying that an IP network is bandwidth efficient but not inherently deterministic. As soon as we set out to impose determinism on the network for example by giving latency sensitive traffic priority, there will be an associated bandwidth and energy cost which includes additional control plane overhead. Most of us experience the impact of high contention ratios over the internet on a daily and hourly basis so it is no surprise that accessing the internet over a mobile broadband network will have similar performance constraints. All that is different is that the bandwidth limits of the physical layer are determined are determined by the amount of available spectrum rather than cable, copper and fibre contention ratios.

Additionally the 5G standards support a number of dual connectivity user cases which could either help or hinder the delivery of deterministic end to end services with defined and closely managed latency parameters.

**3.2 Other Factors influencing latency**

We also need to consider the interrelationship of propagation models, link budgets device performance and latency. There are substantial scatter and absorption losses that need to be accommodated in centimetre band and millimetre band 5G terrestrial networks which need to be characterised in propagation models and channel models. The propagation models determine the link budget and the link budget determines range and throughput and channel coding overhead, often described in the internet world as goodput (the ratio of user bits to coding and control plane bits). However link budgets assume that devices meet a particular conformance standard for example receive sensitivity power output and resilience to unwanted signal energy (dynamic range and ability to manage interference). Theoretically all devices meet their conformance specification but this is verified by measuring devices directly at the output port of the antenna. In practice if the devices are tested in an anechoic chamber, an expensive and time consuming process, they can be shown to perform significantly below the conformance specification, sometimes of the order of 10 dB or more. Conformance specifications may also be relaxed over time, for example if pass bands are made wider or multiple technologies and bands need to be supported in small hand held devices.

Conversely it is possible that devices work better than their conformance specification. An example would be GSM phones which though the 1990’s generally gained about 1 dB per year of sensitivity, with phones at the end of the 1990’s commonly measuring about 7dB above the conformance specification (-102 dBm). This was a consequence of market scale which allowed tighter tolerances to be imposed on RF component supply chains. Sensitivity then steadily worsened as new bands and new technologies (3G and 4G) needed to be supported. Last but not least, devices often fall far short of the performance claimed in the specification sheets because they have been measured in ideal laboratory conditions. Unsurprisingly the end result is that devices work less well than expected in the real world both in terms of their sensitivity, selectivity, stability and output power

The important point to grasp here is that scale helps minimise these implementation issues. More design effort can be applied and supply chains can be bullied to improve raw device performance and the batch to batch and device to device variability of that performance

Last but not least the mechanisms for minimising interference can be influenced by a wide range of internal and external factors. In mobile broadband systems for example including 5G, interference is managed in the frequency and time domain. A TDD network is particularly dependent on maintaining time off sets between interfering devices and accommodating differential delay introduced by flight distance from the device to the base station and multi path. To be efficient, TDD networks coexisting in the same pass band should be clocked together with co sited base stations. This becomes harder to manage as cell size and round trip time increases and is the reason why all satellite networks separate users and channels in the frequency domain rather than the time domain. However there are also issues with higher bit rate high user/device density networks. The latest 802.11ax standard for example introduces FDD into the physical layer as an additional mechanism for managing localised user to user interference but also to accommodate a high density of access points.[[1]](#footnote-1)

TDD timing can be relaxed by increasing the length of the time domain guard either side of a transmitted packet of user data but this absorbs radio network time domain capacity and therefore has a cost. Conversely reducing the time domain guard band requires a more closely toleranced time reference which will have an associated cost.

Last but not least, performance requirements such as operating temperature range can have an indirect impact on latency. Many industrial applications for example are required to work over an extended temperature range which can be as extreme as -40 to +125 degrees centigrade. This places stress on many of the component in the front end of a device including power amplifiers, low noise amplifiers, filters and oscillators. Essentially noise increases with temperature but in the other direction, many components, like us, do not perform well when it gets cold, batteries being a significant example.

Satellite engineers of course have to manage far larger temperature gradients and other pesky issues such as radiation damage and the occasional collision but satellites as we shall show are a critical part of the end to end latency story both in terms of the user experience and routing and backhaul efficiency.

**3.3 Latency and distance and time**

5G and satellite operators have significant ambitions to develop vertical markets where latency is a critical parameter which needs to be managed and controlled. It is important to stress that this is only achievable over short distances. In one microsecond, light and radio waves in free space will have travelled 300 metres so basic physics is going to prevent the delivery of 100 microsecond latency over more than 30 kilometres and this is before you take into account the slower speed of light (and radio) in fibre and routing flexibility.

Figure 3.1 shows the time and distance relationship of radio and light waves in free space

**Table 3.1 Time and Distance for radio and light waves**

|  |  |  |
| --- | --- | --- |
| **Time** | **distance** |  |
| One second | 300,000 kilometres | 186,000 miles |
| One Millisecond | 300 kilometres | 186 miles |
| One Microsecond | 300 metres | 1000 feet |
| One Nanosecond | 30 centimetres | One Foot |

To put this in to a geographic perspective, Singapore is 50 kilometres from east to west and a radio or optical signal will take 166 microseconds to go from one end of this high tech island to the other. Malaysia coast to coast will take one millisecond.

**Figure 3.1 Singapore at light speed**



Australia from the east coast to west coast is 4000 kilometres so that’s a coast to coast travel time of just over 13 milliseconds. Africa North to South is 8000 kilometres so a top to bottom time of 26 milliseconds assuming a direct routing.

End to end latency also depends on the efficient distribution of an accurate time reference across a network particularly if devices are moving in and out of sleep mode to reduce power drain but also poor clocking will increase the likelihood of a loss of synchronisation. The 3GPP vertical market use cases[[2]](#footnote-2) identify a need for more accurate centralized and localized time coordination to support safety critical automotive transport systems, energy grid applications, e-health and m-health and factory of the future applications including a requirement to support end to end latency of less than 5 milliseconds. Requirements for a 5G network described in the IMT2020 Vision September 2015 include a latency of 1 millisecond.

Given that light and RF waves travelling in a straight line take a millisecond to travel 186 miles so by the time transmission loss and routing is added in, 5 milliseconds, let alone one millisecond is an ambitious target over anything other than short distances.

But we have also said that latency and link budgets are closely coupled.

Safety critical vertical markets require a 1 in 105 packet loss threshold. A packet loss threshold is a combination of the packet loss and the end to end latency constraint. Packet loss rates can be reduced by resending packets but this introduces delay and delay variability. A 1 in 105 packet loss threshold might seem a modest target given that fibre is typically specified at 1 in 1012 but cellular networks have typically been designed for 1 in 103 for legacy voice. Moving from 1 in 103 to 1 in 106 requires an extra 3dB of link budget and more closely managed core and edge timing. Every dB of additional link budget translates into a 14% increase in network density. Reducing the packet loss threshold therefore has a direct impact on capital and operational cost.

**3.4 Other Network Overheads and the OSI Model**

There is a saying, probably wrongly attributed to Albert Einstein that the only reason for time is so that everything does not happen at once although at least Einstein understood the significance of this on a cosmic scale. The need to make sure things don’t happen at once is an important aspect of interference management and integration particularly in TDD networks co sharing the same spectrum but also FDD where half duplex is used ( frequency and time domain separation between users). It is also crucial for handover, for channel aggregated multiplexing and for inter-cell interference coordination. Work is ongoing on how to time coordinate 5G with LTE Advanced and LTE Pro co-sharing the same pass band. This will be particularly important for initial non-stand-alone implementations of 5G coupled to the LTE control plane. These same time coordination principles could beneficially be repurposed for 5G and satellite interference coordination.

Satellites are often regarded as introducing long latency but this is an over simplification. End to end delay over a satellite network, particularly a low earth orbit network with inter satellite switching can be quicker than terrestrial in certain conditions. Crucially there is also the second order effect of latency variability also sometimes described as jitter.

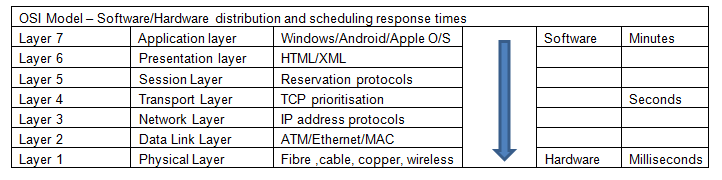
This can be a bigger problem. A known delay can often be accommodated relatively easily but variable delay can be trickier to manage and can unsettle upper layer processes including authentication and end to end security protocols.

This brings us to the protocol stack and the problem of upper layer error control.

In the late 1970’s (the **B**efore **C**ellular era) it was recognized within ISO (the International Organization for Standardization) and the International Telegraph and Telephone Consultative Committee, or CCITT (the abbreviation is from the French version of the name) that there should be a unified standard for describing networking models.

It took a while but in 1984 a unified reference model known as the Open Systems Interconnection Reference Model was published.

**Table 3.2 the OSI model**



This can still today be universally applied to both guided (fibre, cable and copper) and unguided (RF and free space optical) physical layers and is still a convenient and effective way of describing the impact of physical layer (Layer 1) impairments on upper layer performance.

Note that we have added in a rather arbitrary partitioning between hardware and software with hardware still dominant at the physical layer (the low cost power efficient software defined radio is still just round the corner) with the upper layers increasingly implemented as software. As always it is a trade-off between (software) flexibility and (hardware) performance.

From the point of view of 5G and satellite, the point to make is that any physical layer impairments have a multiple cumulative effect at the upper layers of the protocol stack.

A simple example would be automatic repeat requests where the error rate at layer one triggers send again requests. In LTE, these repeat requests can typically introduce up to 8 milliseconds of delay – a combination of delay and delay variability (the delay is an unknown variable). A few automatic repeat requests will trigger upper layer TCP- IP repeat requests and the end result will be reduced throughput, a capacity cost and additional and unnecessary power drain.

**3.5 A Brief History of Time in Mobile Broadband Networks and the impact on latency**

It seems like ancient history now but when GSM was introduced in the early 1990’s it was based on a 20 millisecond frame rate (nicely matched to the voice syllabic rate) supporting a 13 k/bit voice codec with 3 k/bit/second of coding to occupy a 16 kbps channel multiplexed up to the ISDN 144 kbps channel rate. 3G introduced the 10 millisecond time base used in ATM (Asynchronous Transfer Mode) networks. The logic was that 3G networks would need to manage much higher amounts of asynchronous ‘bursty’ traffic and handle different traffic types and traffic priorities in the same channel multiplex.

4G retains the same 10 millisecond time base but introduces sub frames –two half frames- each half frame split into 5 one millisecond frames and LTE Advanced introduces one millisecond as a time base and 5G reduces this to 0.1 millisecond based on the concept of a mini slot. The theoretical benefit is more tight control over layer 1 latency, multiplexing efficiency and power efficiency however the combination of higher data rates and higher level of time resolution requires a more tightly managed and more accurate and stable time base.

Legacy cellular networks such as GSM have relatively straight forward timing and synchronization requirements with frequency synchronization provided via asynchronous Ethernet backhaul using the IEEE 1588 Precision timing protocol and or synchronous Ethernet (Sync E).

Distributed timing using Sync E results in frequency synchronization with an accuracy of 50 parts per billion at the air interface which in turn requires 16 ppb at the base station interface to the backhaul network. The introduction of CDMA in the US introduced an additional need for phase synchronization. This is implemented by using GPS as a frequency and phase reference to an accuracy of between +- 3 to 10 microseconds depending on the cell radius.

In common with CDMA, LTE TDD and LTE Advanced networks also require phase and time synchronization. In frequency synchronized networks, pulse transitions happen at the same rate but not at the same time. They can and probably will have a phase off set. In phase synchronized networks, the leading edge of the pulses occur at an identical moment. In phase and time synchronized networks, the leading edge of the pulses occur at the same time as the phase transition.

The time and phases reference in LTE TDD and LTE Advanced has to be traceable back to Coordinated Universal Time and requires a phase accuracy of +- 1.5 microseconds for cell radii of up to 3 kilometres and +- 5 microseconds for cell radii over 3 kilometres. This is defined by the ITU standard ITU-T G.8272 and needs to compensate for variable delay introduced by router hardware and routing flexibility. The base unit of Coordinated Universal Time is the SI (International System of Units) second. The Si second is defined by a caesium fountain atomic clock.

If you ask a timing expert how accurate a time reference is needed for any given application the answer will always be’ it all depends on….’ One of the dependencies is the time over which the timing accuracy needs to be maintained. For example, a time stamping requirement for a financial transaction or automated computer trading system of less than one millisecond drift compared to universal coordinated time (UTC) can be maintained for three hours independently of GPS using a standard temperature controlled oscillator. Maintaining the same specification over three weeks requires a high specification rubidium source.[[3]](#footnote-3)

Maintaining <1 microsecond of accuracy relative to UTC, needed for example for high frequency trading or smart grid or LTE Advanced mobile networks using a temperature compensated crystal oscillator (TCXO) would support three minutes of holdover. Three hours of holdover would need a highly specified oven controlled oscillator (OCXO) or low specification rubidium source.

Legacy networks have typically been deployed using a master clock frequency specified to the G.811 ITU standard [[4]](#footnote-4) developed to prevent slips in international switch buffers, primarily for speech traffic, but also used as the master clock for systems such as SDH. This has been supplemented by ITU-T G 8272 for time, phase and frequency in packet networks and other recommendations in the G.827x series to compensate for variable delay introduced by switch and router hardware and routing flexibility.

Digital networks since PDH systems in the 1980’s through to SDH and SONET and today’s optical networks have required synchronization. All of these guided media protocols are inherently suitable for synchronization distribution due to the bit-by-bit deterministic way in which they transport data.

The transition to packet networks and Ethernet for backhaul in parallel with the need to maintain legacy TDM networks has meant that synchronization has to be maintained across non-deterministic packet networks.

A common method of achieving this is by using the Precise Timing Protocol (PTP) based on a continuous exchange of time stamped packets which ensures that the grand master clock reference maintains the alignment of boundary and slave clocks. A parallel protocol, the Network Time Protocol (NTP) is used to synchronise computer clocks over a network.

These protocols can be compromised by frame delay (latency), frame delay variation (packet jitter) and frame loss. PTP operates in a similar manner to NTP, but at higher packet rates and generally at the Ethernet Layer rather than the IP layer. This allows PTP to achieve higher levels of accuracy than the one millisecond level generally quoted for NTP systems.[[5]](#footnote-5)

Inconveniently, packet delay in the network is often asymmetric, different between master to slave and slave to master. This complicates the phase synchronization process because the offset computed by the slave will be wrong by the sum of the difference between the two paths.

So for example a computer or server exchanging time stamps every second between a slave and master with 50 nanosecond accuracy could be transitioning through a switch or router introducing asymmetric path delay (packet delay variation) of the order of tens of microseconds.

The computer or server will be running an operating system coupled to a quartz oscillator which can add microseconds of error per day and there will be an additional difference of several microseconds depending on whether the server is loaded (with the fan running) or unloaded. The filling and emptying of traffic buffers causes additional asymmetric delay variation.

The impact of this is that the core network reference has to be at least an order of magnitude more accurate than the boundary clock reference, for example one microsecond at the edge will need 100 nanoseconds at the core. This level of accuracy is also needed to provide back up when GPS is unavailable.

There seems to be an emerging consensus within the 5G standards community that there will need to be a reference time accuracy at the network edge of the order of 300 to 500 nanoseconds which implies 30 nanoseconds at the core though it is hard to see how useful this will be if the other causes of end to end delay cannot be measured and managed and could potentially result in unexpected edge timing and synchronisation costs.

This also implies a need to qualify the timing needs of network function virtualization (NFV), assumed as one of the prime mechanisms for reducing delivery cost in 5G networks - a badly timed virtual network will by implication be a badly behaved virtual network. Packet timing protocols work adequately well over Layer 2 (the data link layer) but not Layer 3 (the network layer) and expensive work arounds may be needed which will negate the promised cost benefits.

The default answer is to use GPS with the comfort and assurance that GPS is becoming more accurate and resilient to jamming with the addition of the L2 and L 5 frequencies, launch of the Galileo and Beidou constellations and enhanced upgraded Glonass but getting GNSS signals into buildings can be hard and expensive. Lightning strikes or high winds can take out external antennas and satellite signals are subject to space weather effects. A wired alternative therefore continues to be a desirable back up. Some countries are investing in additional time reference systems that can act as an additional back up to GPS. The Quazi Zenith constellation being implemented in Japan is one example. [[6]](#footnote-6)

Generally it can be stated that as bit rates increase and the number of users and access points increase and as more networks are locked together in the time domain with time domain mechanisms used to manage interference it will become increasingly important to maintain and distribute an accurate clock reference.

**3.6 The Cost of Accuracy**

A possible longer term alternative is to have very accurate clocks distributed through the network and in edge devices. There are emerging low cost atomic clock options developed originally by DARPA to provide accurate dead reckoning for missiles flying when their GPS reference is jammed. The devices are known as Chip Scale Atomic Clocks. The principle of miniature atomic clocks is based on a technique known as Coherent Population Trapping using a compact sealed vacuum cell of a few cubic millimetres which contains an alkali vapour which is illuminated by a high frequency modulated laser beam. A device available today produced by Symmetricom, uses caesium 133 and a buffer gas in the resonance cell. The vapour is illuminated with a semiconductor laser modulated at a frequency close to the natural oscillation frequency of the caesium atoms, about 9.192 GHz. As the caesium atoms start to oscillate, they absorb less light and the photons transmitted through the cell are used to determine when the modulation frequency of the laser beam coincides with the resonant frequency of the atoms. It is effectively an atomic phase lock loop. The Symmetricom clock weighs 35 grams and draws 115 milliwatts of power and measures 4 by 3.5 by 1.1 centimetres. It is accurate to within less than half a microsecond a day and can work across a -10 to +70 degree Celsius operating range.

This makes it useful for a whole range of applications including back pack military radios, military GPS receivers, unmanned aerial vehicles, back pack IED jammers and marine geophysical sensors (GPS doesn’t work under the sea!). At around $1500 dollars it has not yet achieved consumer price levels but as prices fall and accuracy improves these miniature clocks will become significantly useful in 5G mobile broadband and telecommunications timing and positioning systems. As with all electrical equipment, these devices will however be subject to electrical failure.

Improving the accuracy of grand master clocks is also both desirable and necessary for 5G but has cost implications. An optimised cesium clock costs around $100,000 dollars but cesium depletion means that the cesium tube will need replacing somewhere between every five and ten years at a cost of $30,000 dollars.

Strontium based atomic clocks are being suggested as an alternative [[7]](#footnote-7) as are optical clocks [[8]](#footnote-8) but cesium and rubidium based devices remain as the default sources of accurate time in present and future networks for at least the next few years.

However there are other performance parameters including start up time that are critical to radio network applications including broadcasting, satellite and terrestrial mobile broadband which introduce additional synchronization cost. The better clocks typically have much longer start up and stabilisation times.

Resilience is also a cost and generally dependent on supplying multiple clock sources. The repurposing and recommissioning of legacy Loran Very Low Frequency (VLF) transmitters is being studied and tested as a cost effective way to provide UTC traceable time to applications in GNSS denied environments. Initial test programme results suggest this could yield UTC traceable results with an accuracy of better than 100 ns, a quality comparable to GPS but with better indoor penetration.[[9]](#footnote-9) Supplementary system innovations such as e-Loran are therefore useful as potential additional time sources.

**3.7 Time, latency and Network Function Virtualisation**

We have said that the transition from 4G to 5G implies a need for higher data rates, lower end to end latency, better resiliency, lower packet loss thresholds and low packet delay variability. These together with advanced interference management techniques in the radio layer imply at least an order of magnitude improvement in time accuracy both at the core and in boundary clock devices.

This improvement will also be needed to support network function virtualization (NFV). In particular the promised cost efficiency gains of NFV may be at least partially offset by additional synchronization costs. At the very least it is to be expected that synchronization costs are likely to increase as a percentage of network deployment costs as we move from 4G to 5G networks.

Clock quality is equally critical to all guided media including next generation cable (DOCSIS 3), copper (G.fast) and fibre (GPON) as is the time domain integration of the radio access layer with copper, cable and fibre back haul.

It is plausible to claim and probably possible to prove that clock quality value, the difference between the cost of improving clock quality versus the additional realized value at network and device level increases as bit rate increases. Traffic per watt efficiency will also increase although it is hard to find a costed analysis of this.

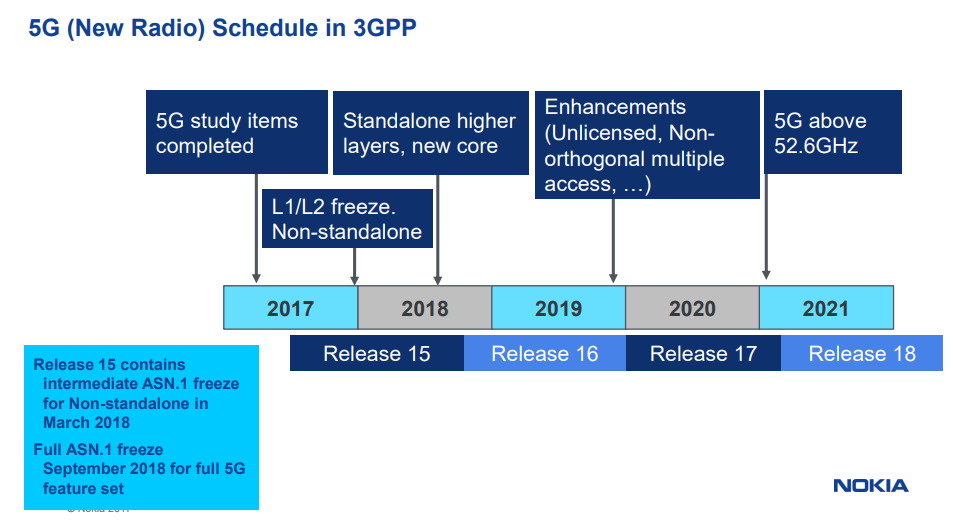
**3.8 New Radio Specification and related latency issues**

The sheer complexity of the 5G standards process can make the analysis of a single aspect, in this case latency, really quite difficult but let’s apply our best effort to the process.

At the time of writing of this book, the freeze date for Release 15 is set at September 2018 with the Release specifying eMBB and some parts of URLLC. There is however a sub set of Release 15 which is scheduled for Freeze six months earlier known as Non Stand Alone where 5G is coupled closely to the time base and frame structure and control plane topology of LTE. Non standalone includes changes to the core network designed to make it more flexible when managing multiple traffic multiplexes with many different latency and throughput requirements.

Figure 3.2 summarizes the time lines which include work items on 5G in unlicensed spectrum and the proposed schedule for work on 5G above 52.6 GHz (1 and 2 GHz channel bandwidth 5G).

**Figure 3.2 5G New Radio Schedule With thanks to Nokia Networks**[[10]](#footnote-10)



The specification includes quality of service (QOS) and policy frameworks both of which will have a direct impact on the latency delivered to individual users and devices, network sharing (which theoretically at least should improve multiplexing efficiency) and trust hierarchies. Incidentally authentication protocols can have a major impact on latency as they can introduce milliseconds of delay at a local level and seconds of delay at network level. We have also made the point earlier that latency and delay variability can compromise the authentication process more or less ensuring that a session never gets started.

At the physical layer, the big time domain difference at frame level is the introduction of the mini slot with the specific objective of making present and future URLLC requirements easier to meet. For example a mini slot can be used for URLLC pre-emption in an eMBB shared channel. The mini slot is also intended to help operation in unlicensed bands by allowing transmission to start directly after a successful listen before talk procedure without waiting for the slot boundary.

Mini slots consist of two or more symbols – the first symbol includes uplink or downlink control information. Figure 3.2 shows how a 10 millisecond radio frame is divided down.

**Figure 3.2 Mini Slots in 5G- With thanks to Nokia Networks**

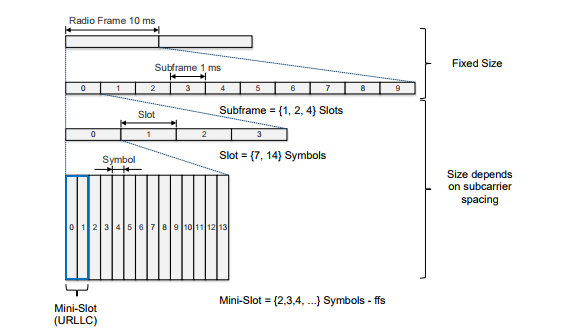
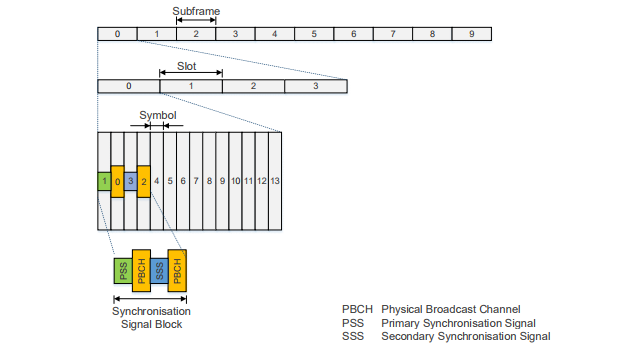


Figure 3.3 summarises the sub frame structure and synchronisation signal blocks

**Figure 3.3 5G sub frame structure and synchronisation signal blocks - with thanks to Nokia Networks**



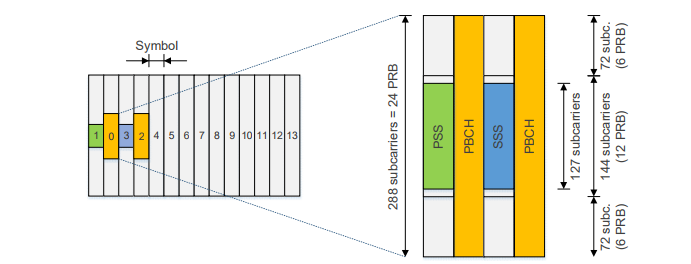
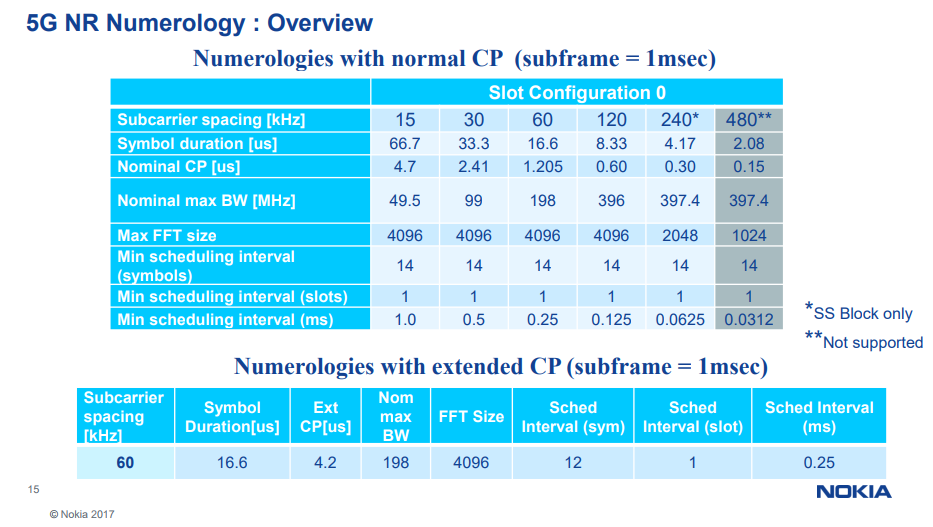


Figure 3.4 shows how the minimum scheduling intervals scale as the sub carrier sub spacing increases from one miillisecond with 15 KHz sub carrier spacing to below 0.1 millisecond for 240 KHz and 480 KHz sub carrier spacing. An extended cyclic prefix option is also shown (for larger cells or less closely managed timing – see earlier comments) with a scheduled interval of 0.25 milliseconds.

**Figure 3.4 Minimum Scheduling Intervals Scaled against Sub Carrier Spacing - With thanks to Nokia Networks**



The TDD frame structure is slightly different and divides 20 millisecond frames into two 10 millisecond sub frames. The example shown in Figure 3.5 is for a 120 KHz sub carrier implementation. Note that one advantage of TDD is that the uplink and downlink are reciprocal (on the same centre frequency) which makes channel sounding simpler which in turn makes it easier to beam form. The graphic also shows the signalling overhead embedded within the frame structure.

**Figure 3.5 TDD Frame Structure- With thanks to Nokia Networks**

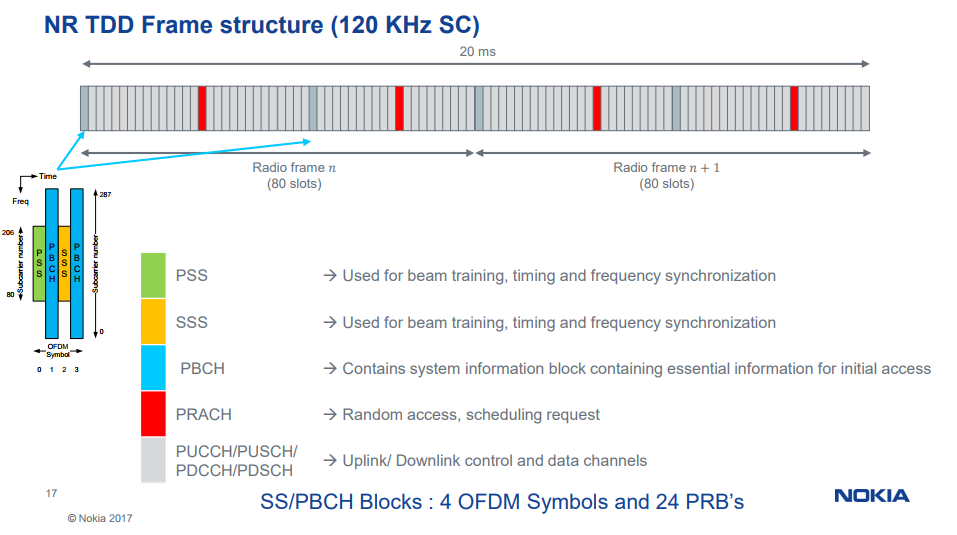
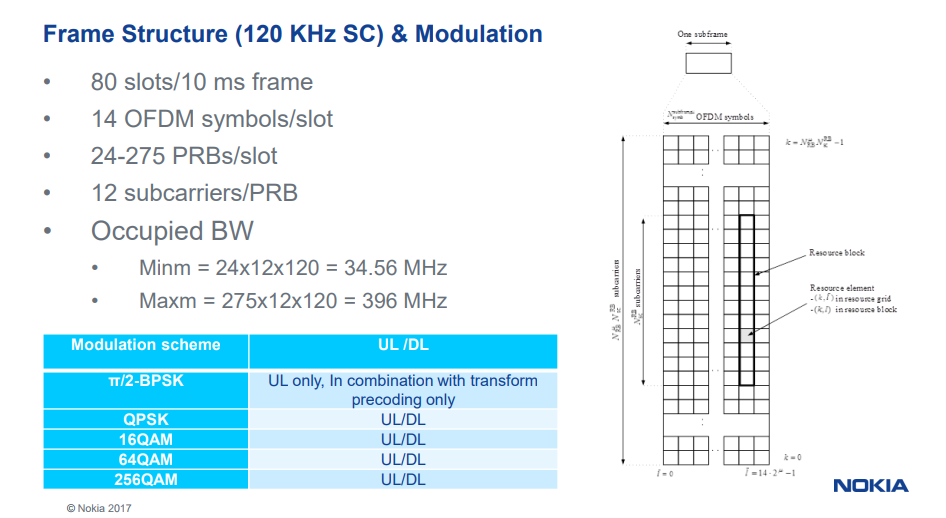


Figure 3.6 shows the uplink modulation options from two level PSK for power constrained devices to 256 QAM for high throughput.

**Figure 3.6 Modulation Options- With thanks to Nokia Networks**



**Figures 3.7 and 3.8 show the coding schemes applied to the data channel and control channel**

Note that the control channel will generally be more heavily coded than a data channel. The modulation and coding used on the data channel in particular will adapt to changing channel conditions to minimise the triggering of physical layer and higher layer send again messages.

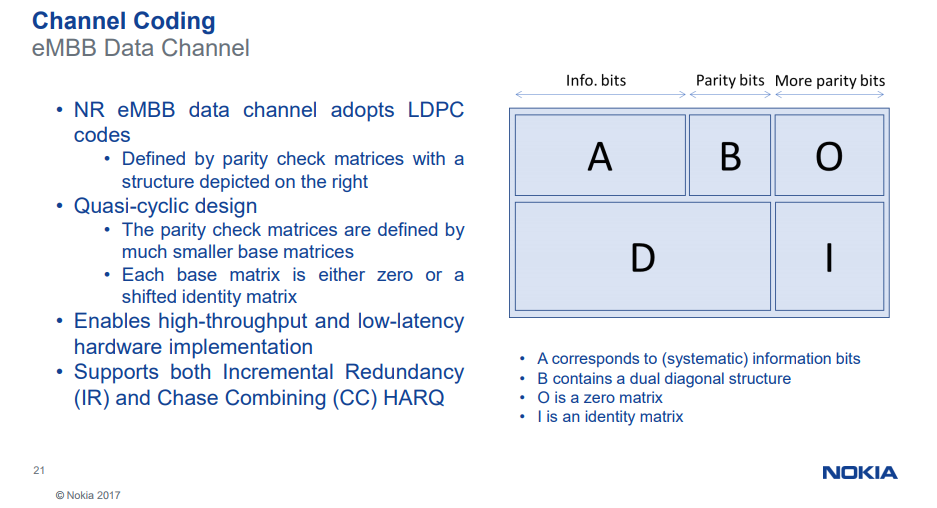
The accuracy and speed with which a network is able to make channel quality measurements is therefore another factor with an impact on end to end latency.

Much work continues to be invested in optimising channel coding to improve error detection and correction. The end objective is to minimise residual bit error rate as the signal to noise and carrier to interference conditions become more adverse. Generally all schemes combine some block coding with parity checks and convolutional coding. Block coding introduces some delay and convolutional coding (coding with memory) introduces some delay variability.

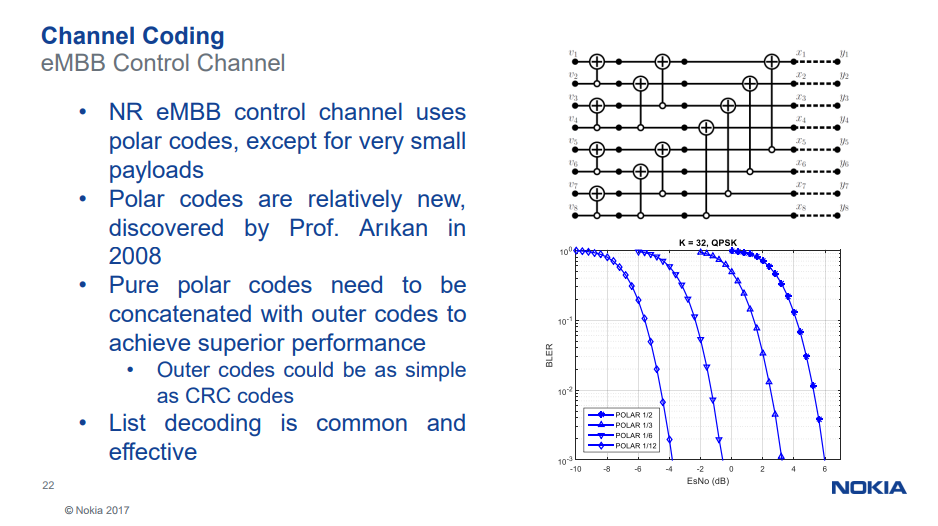
The reason the parity checks referenced in Figure 3.7 are short is to minimise the impact on physical layer latency, techniques such as chase combining[[11]](#footnote-11) are used to limit the number of clock cycles and time needed to error detect and correct a wrongly decoded symbol.

To understand polar codes really requires a higher degree in maths but if you feel up to it, follow the link for more information.[[12]](#footnote-12)

**Figure 3.7 Channel Coding on an eMBB Data Channel- With thanks to Nokia Networks**



**Figure 3.8 Channel coding on an eMBB Control Channel­- With thanks to Nokia Networks**



In an ideal world signal to noise levels and carrier to interference would be maintained at a level that would minimise bit errors on the radio path. This is relatively easy to achieve in guided media such as fibre, cable or copper because the impairments are predictable and stable and can therefore be managed and mitigated. This is why a fibre physical layer can be held at a 1 in 1012 for example. Unguided wireless links particularly in mobile networks are harder to manage because impairments such as multipath phase cancellation change rapidly and unpredictably. Channel coding in mobile networks allows user bit error rates to be minimized in poor signal to noise or carrier to interference conditions but there is a bandwidth cost (the addition of extra error correction bits) and latency cost (send again instructions when error rates exceed a certain threshold).

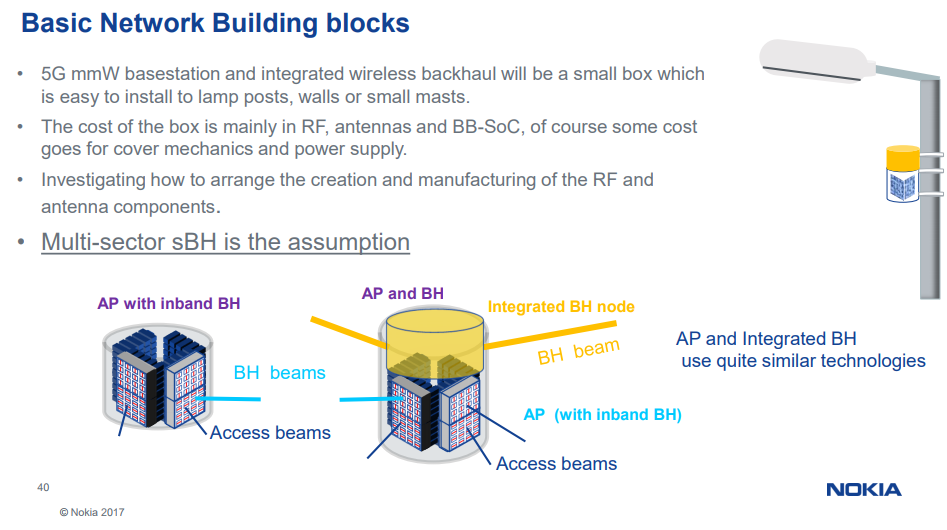
**3.9 In Band Backhaul**

Having temporarily exhausted the topic of 5G physical layer frame structures and channel coding (we come back to the topic in later chapters), it is time to take a look at terrestrial backhaul and its impact on end to end latency

Backhaul can either be via microwave link or fixed copper or fibre or free space optical or if none of these options is cost effective then via satellite.

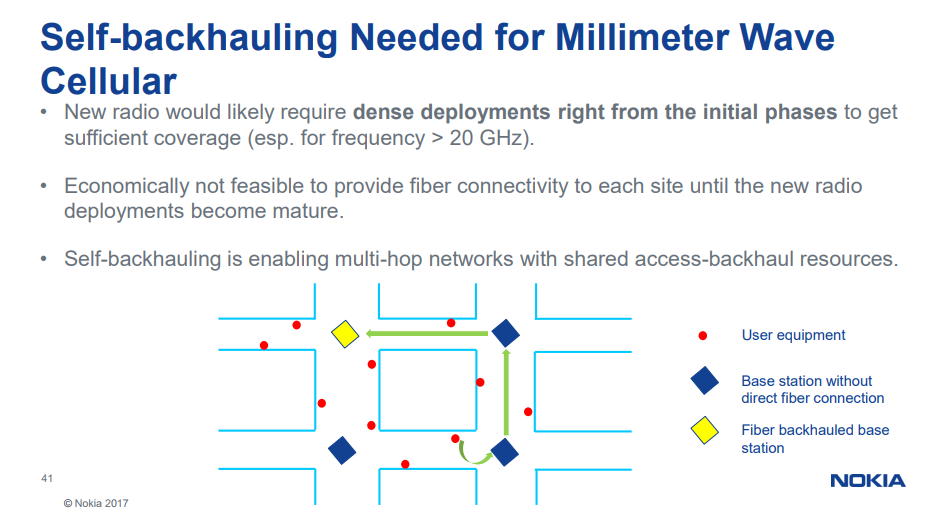
The difference in 5G is that it makes a lot of sense to use the same pass band to support users and backhaul. This is known generically as in-band backhaul or self-back hauling and has a number of advantages including the opportunity to reuse RF hardware and baseband processing. It also avoids the need to demodulate and modulate and channel code traffic as it moves from the radio layer into the backhaul network. Note that backhaul in existing networks is often at 28 GHz or 39 GHz so the RF hardware already exists. Given that satellites are also using this spectrum then there is an obvious opportunity to intensively reuse the spectrum with the important caveat that cross system interference has to be rigorously managed. It is an ongoing narrative in this book that this process is critically dependent on achieving effective spatial separation between potentially interfering systems and for users of those systems to be confident that these techniques can be made to work across multiple systems administered and managed by multiple operators.

**Figure 3.9 In Band Backhaul in 5G** **- With thanks to Nokia Networks**



Nokia make the obvious but important point that in a dense or ultra-dense network it is unlikely that fibre would be economic at least for initial roll out. The self-back hauling is used to get to the nearest fibre end point.

**Figure 3.10 In Band Self Backhaul in 5G** **- With thanks to Nokia Networks**



If well designed and implemented, this topology should not add materially to the overall latency budget.

Overall it can be seen that much effort is being invested in standardising the 5G physical layer and the supporting network topology in order to meet assumed 5G end user end to end latency and throughput requirements

This work and the calculated throughput and latency metrics achievable provide a benchmark against which GSO and NEWLEO operators will need to measure their constellation performance.

This brings us logically to a comparison of 5G terrestrial channel modelling and propagation and satellite channel modelling and propagation. Superficially it might be considered that if a terrestrial system and satellite system are implemented in the same pass band, the two systems will have similar propagation constraints. If implemented with similar channel bandwidth and channel spacing then it might be thought that would have similar channel characteristics.

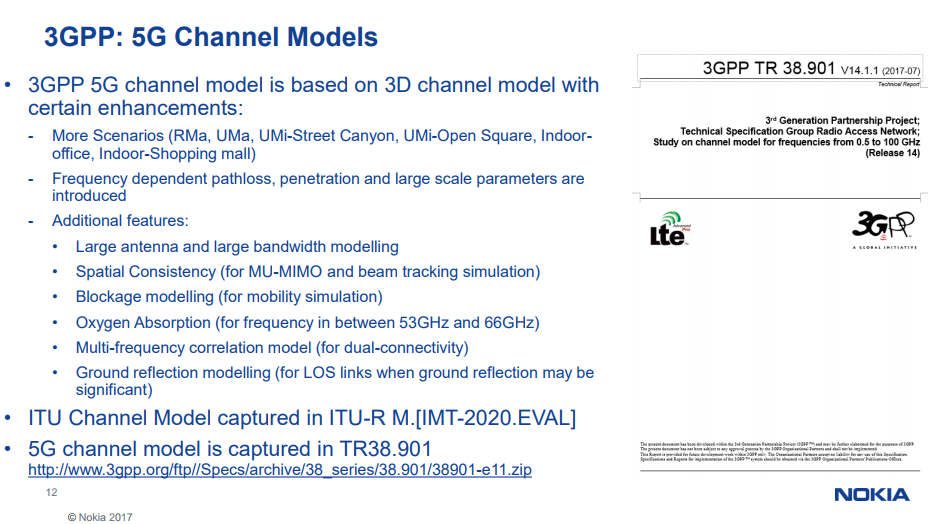
However the terrestrial user to 5G base station path is different to the terrestrial user to space path. Additionally different satellite constellation topologies will have different propagation properties but as we shall see, the angle of arrival and angle of departure of all systems both terrestrial and space based ultimately determines throughput, end to end latency and carrier to interference coexistence.

**3.10 5G channel models**

The starting point in this discussion is the 5G channel model.

These are described in the 3GPP document TR 38.901 but obliging those nice people at Nokia Networks have summarised the main work items and associated channel model references.

**Figure 3.11 5G Channel Models - With thanks to Nokia Networks**



If you wanted a one line summary of all of this then it would be reasonable to say that 5G in the centimetre and millimetre band works very much better when line of sight but unfortunately this is quite a rare occurrence in urban and rural topographies both for mobile and fixed users and even line of site from a low elevation, for example a base station a few metres from the ground can be subject to substantial non-additive ground reflection. This is the single most important advantage that high count LEO constellations in the K bands or V and W band/E band have over terrestrial 5G because (as we keep saying) these constellations will be nearly always nearly overhead. If 10,000 LEO satellites get launched into space then there will be a satellite more or less directly overhead more or less all of the time more or less anywhere in the world.

This not only minimises the earth to space latency budget but minimises the path through the atmosphere. The signal will also be minimally affected by surface scatter or ground reflection.

Frustratingly it is presently hard to quantify exactly how much power from a terrestrial RF transmitter actually gets lost on its way to its local destination.

Contemporary cellular networks at UHF or L band and C band are designed using well established propagation models with physical layer RF and baseband parameters determined by a range of user defined pedestrian ‘typically urban’ (TU3), vehicular urban (TU50) and rural channel models (RA250). These work adequately well up to 4 GHz but become progressively less accurate at higher frequencies/shorter wavelengths.

Discussions around suitable channel models for the centimetre and millimetre bands focus on the relative merits of the Alpha Beta Gamma (ABG) model using a floating constant referenced to known and measured data sets, the Close In (CI) model referenced against a path distance of one metre and a CIF model which adds in a frequency weighted path loss exponent but no large scale existing data sets presently available to verify/fine tune these models. Anecdotally the observation is made that the ABG model typically under predicts path loss when near to the transmitter and over predicts path loss further way. The CI and CIF models are more accurate and computationally simpler. The CI model works better for outdoors and CIF model for indoors. Both models have a path loss variable that is continuously coupled to the transmitted power over distance.

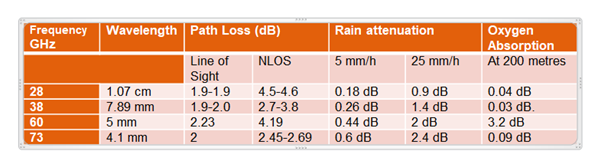
Within the EU, research work has been focused on frequencies from 2 GHz to 73 GHz with a path length of between 4 metres and 1238 metres with models for urban microcells (UMi) with antennas at 10 metres, urban macro cells (Uma) with antennas at 25 metres and Indoor Hot Spots (InH).

The measurements and modelling are based on narrow beam 7.8 degree azimuth half power beam width antennas and wideband 49.4 degree antennas. While this work will almost certainly yield useful outputs it does not include modelling for larger cells. Neither does it set out to model fractional beam width antennas (half power beam width of between 0.5 and 1.5 degrees)

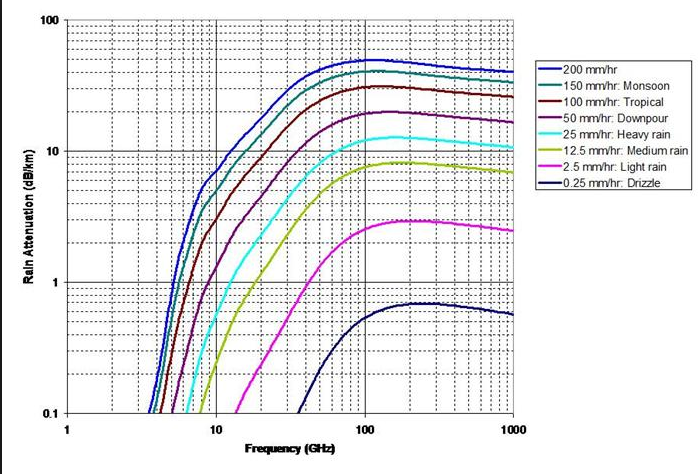
Fortuitously substantial modelling does exist for point to point backhaul which yields simple but useful path loss estimates for specific frequencies for line of sight and non-line of sight links in a range of atmospheric conditions (rainfall rates) and taking into account oxygen absorption (peaking at the oxygen resonance frequency at 60 GHz). Note how the wavelengths above 30 GHz to 60 GHz (10mm to 5mm) are similar to or less than the roughness of many man-made and natural surfaces hence the high absorption and surface scatter.

If you wanted a one line summary of this graphic then it would be that for a 28 GHz receiver needs an additional 30 dB of isotropic gain to see the same amount of power as a 900 MHz receiver. As we shall see in Chapter 6 this is not that hard to achieve. To restate, the main enemy here for 5G terrestrial is surface scatter and absorption and non-line of sight losses.

**Table 3.3 Centimetre and Millimetre Band Propagation Measurements and Modelling used to design wireless backhaul**

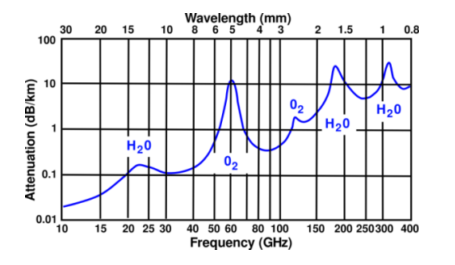
****

|  |  |
| --- | --- |
| |  | | --- | |  |   Figure 3.12 reproduces the ITU rain fade models  **Figure 3.12 ITU Rain Fade Models** |
|  |



For the sake of completeness, here is the much reproduced graphic showing the water and oxygen resonance lines.

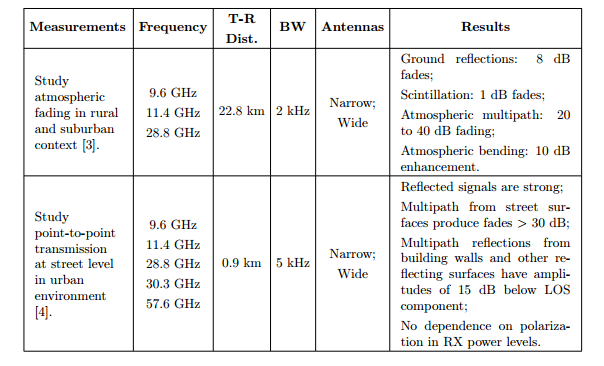
**Figure 3.13 Water and Oxygen Resonance Lines – with thanks originally to the Rutherford Appleton Laboratory**



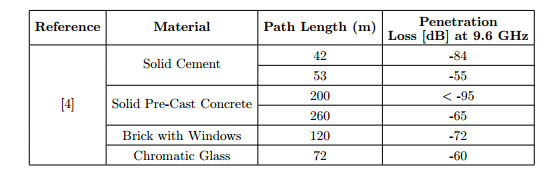
An additional issue is that channel measurements need to be made across a wide range of channel bandwidths from a few KHz to 1 or 2 GHz.

The additional cost of wider band spectral analysis means that there tend to be more narrowband measurements available. The following tables show a range of narrow band measurements made by the wireless research team at New York University.

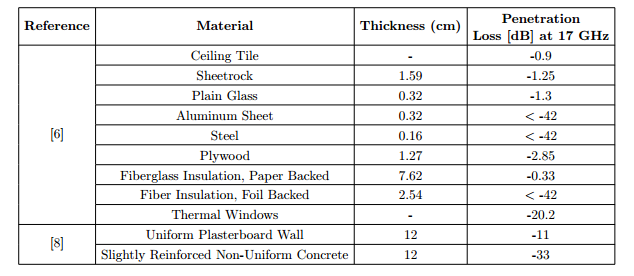
**Table 3.4 Narrow Band Measurements made by New York University[[13]](#footnote-13)**



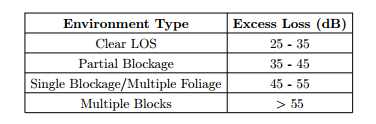
**Table 3.5 Comparison of the penetration loss of a range of materials**



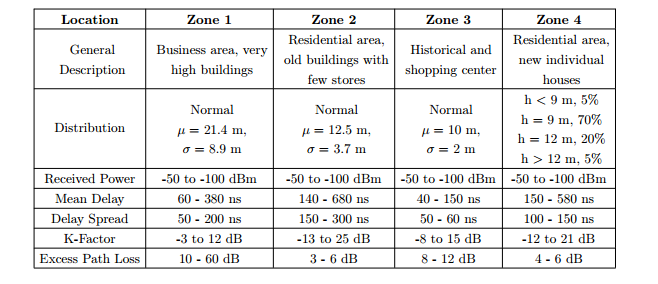
**Table 3.6 More penetration loss numbers**



**Table 3.7 Blocking including foliage (will be higher when wet)**



**Table 3.8 Delay spread in various environments**



Incidentally the table indicates a large variation in K factor. K factor was discovered by British radar engineers working around 300 MHz in the 1930’s and was the factor used to dimension the observed phenomena that radio waves followed the curvature of the earth and therefore travelled further than the horizon limit. The bending is proportional to frequency, with waves bending less as the frequency increases. This curvature propagation effect was calculated as being equivalent to increasing the diameter of the earth by 33% though it can also be affected by atmospheric conditions which also have a progressively larger impact at higher frequencies.

**3.11 Satellite Channel Models and signal latency**

So the story so far is that propagation conditions become more variable as frequency increases, propagation losses are higher but other factors such as surface absorption and scatter and atmospheric conditions also become increasingly important. Networks need to work in the rain.

These factors are a dominant influence in deciding which of the satellite constellation options could be best suited to complement 5G networks. The options include geostationary satellites, medium earth orbit satellites and low count and high count low earth orbit constellations.

In all cases the channel models and propagation characteristics will be determined by the elevation angle as seen from earth

Self-evidently a terrestrial device looking at a GSO satellite over the equator will be subject to increasing propagation loss at higher latitudes. As latitude increases, the path length gets longer and the signal will pass through more of the atmosphere. The coverage pattern from the satellite will be roughly circular over the equator and increasingly elliptical at higher latitudes so the flux density will be lower at higher latitudes. The longer path length will also increase end to end delay.

Low count MEO and LEO constellations will also be serving users on earth at a variety of inclination angles depending on their position in the sky at any given moment. As we have said several times already, high count LEO and MEO constellations will be nearly always overhead nearly all the time.

The speed of light in free space is more or less constant bar some gravitational effects so flight distance and flight time for all satellites in all orbits can be precisely calculated and are determined by orbital altitude and elevation angle.

**GSO satellites** orbit the earth at 36,000 kilometres above the equator. Radio waves go at the speed of light which is 300,000 km per second. For users on the equator communicating with a satellite directly overhead, the total distance, single hop (up and down) is 72,000 km so the time delay is 480 ms for a round trip.

A geostationary satellite is visible from a little less than one third of the earth's surface and if you are located at the edge of this area the satellite appears to be just above the horizon. The distance to the satellite is greater and for earth stations at the extreme edge of the coverage area, the distance to the satellite is approximately 41,756 km. Communicating with another similarly located site is a distance of nearly 84,000 km so the end to end delay is almost 280 ms (one way). Extra delays occur due to the length of cable extensions at either end, and if signals are routed by more than one satellite hop. Significant delay can also be added in routers, switches and signal processing points along the route.

In a **MEO network** (using O3b as an example) orbit height is 8,062 km. A typical single hop path involves sloping path lengths of 11,000km producing a single hop distance of 22,000 km producing a latency of 73 milliseconds. O3B claim[[14]](#footnote-14) a round trip latency of better than 150ms based on a double hop distance of 11,250 + 11,250 + 11,250 + 11,250 km.

In **LEO networks** the propagation delay is smaller still. Iridium’s constellation operates at 780 kilometres, Orbcomm is a little higher at 825 kilometres and Globalstar is at 1,414 kilometres. The propagation delay experienced in a LEO satellite system varies as the satellites change position but will be 4.3 milliseconds per hop for Iridium, 4.5 milliseconds for Orbcomm and 7.8 milliseconds for Globalstar for ‘bent pipe’[[15]](#footnote-15) applications with the satellite directly overhead. These figures should be doubled for round trip delay.

If the terrestrial end points are not within the coverage of a single satellite (this varies with each system) then the distance will be greater, with inter-satellite links via other satellites.

Propagation delay is however only one part of the delay budget. Delay and delay variability is also introduced by processing delay for example through any router nodes or relay transponders. If these devices are software configurable then delay variability could be significant. This is sometimes described as serialization delay. The delay through relay transponders is a function of forward error correction and modulation.

Satellite systems over the past twenty years have evolved from initially using Viterbi coding then Reed Solomon coding then turbo codes (codes with memory). As data rates increase, block sizes increase and convolutional coding delivers more coding gain and avoids ‘send again’ loops which introduce delay variability. Satellite TV provides a noticeable example of propagation and coding delay when compared to terrestrial TV.

The trend in satellite schemes below 2 Mbps is to use smaller forward error correction block sizes. These can reduce a typical 200 millisecond error correction induced delay to 50 milliseconds. In other words 300 milliseconds can be eradicated from the round trip delay by changing the forward error correction scheme.

Last but not least TCP/IP uses acknowledgements to determine the amount of bandwidth available before any data is exchanged between two points (the slow start algorithm). This requires three round trips or six satellites hops just to get started and if the session goes idle the whole process has to start again. There are various ways round this including TCP/IP Fast Start and caching and local storage in off peak periods. These approaches are now standardised.[[16]](#footnote-16) There are also a number of WAN optimizer and WAN accelerators from a cross section of vendors typically implemented at Layer 4 of the OSI protocol stack.

On the one hand this would suggest that satellites introduce substantial additional delay over and above terrestrial networks.

On the other hand a satellite constellation provides absolute visibility across the whole end to end channel which means that delay and delay variability can be calculated and compensated. This is harder to achieve in terrestrial networks.

This may seem trivial but is actually an important point. Often it is not the actual delay that is the problem but rather **the second order effect of delay variability**. It is hard to design challenge and response and authentication or send again algorithms when these variables vary over time.

A simple example is me asking you if are you called John.

I expect an answer within a certain elapsed time. If you fail to answer for half an hour I would have a nagging doubt that either your name was not John or if it was you didn’t want to admit it.

In a communications network, authentication challenge and response algorithms have specific expectations of how long each part of the authentication should take. This is a simple but effective way of reducing the chance that the algorithm has been spoofed. A satellite link with a known end to end delay and known and calculable delay variability should be more secure than a flexibly routed exchange across multiple terrestrial networks..

**3.12 Ongoing Satellite standards and related study items**

It has taken a while but as we started writing this book (September 2017) a standards group studying the potential touch points between 5G new radio standards and non-terrestrial networks finally ground into action.

The work group has a rather narrow group of companies involved[[17]](#footnote-17) (and no obvious participation from the mobile broadband terrestrial standards community) but at least it is a start.

The3GPP TR 38.811 2017-06 study [[18]](#footnote-18) on New Radio to support Non Terrestrial Networks covers bent pipe payloads and regenerative satellite topologies which is satellite speak for relays and repeaters. A bent pipe takes a signal from earth, amplifies it and sends it back again. A regenerative transceiver demodulates and decodes and then modulates and recodes the downlink

A regenerative transceiver on a satellite therefore introduces additional processing delay though this is immaterial when compared to the round trip delay. The advantage is that the signal is cleaned up prior to its onward journey.

The bent pipe is therefore effectively performing the same function as an LTE repeater. A regenerative transceiver is analogous to an LTE relay.

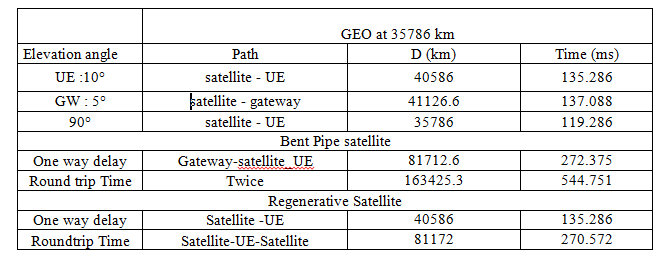
**Figure 3.13 Bent pipes, regenerative transceivers versus terrestrial repeaters and relays**

|  |  |
| --- | --- |
| LTE Repeater | LTE Relay |
|  |  |

In a bent pipe one way propagation delay is the sum of feeder link propagation delay and user link propagation delay. A regenerative payload is essentially similar but with a bit of on board processing delay added in.

The document includes a table showing the propagation delay as a function of path distance from GSO satellites at different elevation angles

**Table 3.9 GSO Propagation delay as a function of path distance**



Similar metrics are then presented for Non-Geostationary (NGSO) LEO and MEO satellites

**Table 3.9 NGSO Propagation delay as a function of path distance**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  |  | LEO at 600 km | | LEO at 1500 km | | MEO at 10000 km | |
| Elevation angle | Path | Distance D (km) | Delay (ms) | Distance D (km) | Delay (ms) | Distance D (km) | Delay (ms) |
| UE: 10° | satellite - UE | 1932.24 | 6,440 | 3647.5 | 12,158 | 14018.16 | 46.727 |
| GW: 5° | satellite - gateway | 2329.01 | 7.763 | 4101.6 | 13.672 | 14539.4 | 48.464 |
| 90° | satellite - UE | 600 | 2 | 1500 | 5 | 10000 | 33.333 |
| Bent pipe satellite | | | | | | | |
| One way delay | Gateway-satellite\_UE | 4261.2 | 14.204 | 7749.2 | 25.83 | 28557.6 | 95.192 |
| Round Trip Delay | Twice | 8522.5 | 28.408 | 15498.4 | 51.661 | 57115.2 | 190.38 |
| Regenerative satellite | | | | | | | |
| One way delay | Satellite -UE | 1932.24 | 6.44 | 3647.5 | 12.16 | 14018.16 | 46.73 |
| Round Trip Delay | Satellite-UE-Satellite | 3864.48 | 12.88 | 7295 | 24.32 | 28036.32 | 93.45 |

**3.13 The impact of NEWLEO progressive pitch on latency and link budgets**

The descriptive work items do not presently include any modelling of the impact of progressive pitch where the elevation angles are altered to minimise interference into GSO and MEO ground base gateways and user terminals. This is probably due to the present lack of engagement in the standards work by any of the NEWLEO operators but will need to be addressed at some stage, sooner rather than later.

**3.14 Summary**

Many factors determine latency and delay variability in 5G terrestrial and satellite networks including link budgets and path length. Counter intuitively satellites can have a link budget advantage provided they are almost directly overhead, for example for GSO satellites over the equator or for high count LEO and MEO constellations at any latitude.

Nearly always nearly overhead or ideally always overhead will generally give a good line of sight into most outdoor coverage scenarios and surface absorption and ground reflection will be minimal.

For the same reason, satellites are not effective at providing connectivity to hand held devices in buildings due to the high penetration losses of typical building structures. The only time satellites would have visibility inside a building is if they were visible at a low inclination angle and terrestrial networks would probably be more effective in these conditions. Satellite operators overcome this constraint by supplying antennas to mount on window ledges or building roofs. VSAT terminals for corporate connectivity for example have been used for many years but the shift is now towards low cost transceivers with integrated Wi-Fi for local in building coverage. The OneWeb model for low cost connectivity outdoors can reuse this in builsing hardware with a Ka band transceiver coupled to Wi-Fi. This means that Wi-Fi will add some delay to the latency budget though this should not be material in the context of the overall round trip delay.

Although round trip delay might seem to eliminate satellites from many 5G latency critical use cases, there are some circumstances where satellites can deliver a performance advantage. Examples already referenced in earlier Chapters include Iridium and potentially LEOSAT where the faster speed of light and radio in free space (compared to fibre) outweighs the path delay over distances of more than 10,000 kilometres. Iridium also has inter satellite switching (K band) and LEOSAT and Space X propose to use optical inter satellite switching all of which provide absolute control of the end to end channel this eliminating the second order effect of delay variability. This can be particularly useful if very high level authentication is needed. This combination of end to end control and security is a basic requirement for military and life critical systems but also for high value financial services such as high frequency trading.

Latency however is also a function of the loading imposed on a network. This is known as ‘fill factor’ in the satellite industry but can also be expressed in all wireless and wireline networks as ‘contention ratio’ which is a function of the number of users co sharing a radio channel, fibre, cable, copper resource or router node. The number of users and the burstiness of the aggregated offered traffic from these users define the required delivery and buffer bandwidth of the network. The buffers are provisioned to allow for queuing so that packets do not get lost or discarded when there are insufficient delivery resources at any point in the end to end channel. In an ideal world a network would be dimensioned to avoid queuing. This would also avoid the need for traffic prioritisation and pre-emption, the process of separating out latency tolerant and latency sensitive traffic which in itself absorbs bandwidth and power. This would however involve dimensioning the network to accommodate worst case loading which would be expensive both in terms of capital and operational cost. Latency is therefore partly a product of network topology but also a product of network economics.

Bearing all this in mind, the question to answer is whether satellites can deliver services to earth at an equivalent or lower cost than terrestrial 5G.

The next six chapters set out to find the answer to this question.

1. See Chapter 10 for more detail [↑](#footnote-ref-1)
2. http://www.3gpp.org/news-events/3gpp-news/1839-5g\_cc\_automation [↑](#footnote-ref-2)
3. http://www.chronos.co.uk/index.php/en/time-timing-phase-monitoring-systems [↑](#footnote-ref-3)
4. https://www.itu.int/rec/T-REC-G.811/en [↑](#footnote-ref-4)
5. http://www.ntp.org/ [↑](#footnote-ref-5)
6. http://qzss.go.jp/en/ [↑](#footnote-ref-6)
7. http://www.nist.gov/pml/div689/20150421\_strontium\_clock.cfm [↑](#footnote-ref-7)
8. https://www.rp-photonics.com/optical\_clocks.html [↑](#footnote-ref-8)
9. http://www.chronos.co.uk/index.php/en/delivering-a-national-timescale-using-eloran [↑](#footnote-ref-9)
10. Dr Amitabha Ghosh IEEE Webinar September 20th 2017 [↑](#footnote-ref-10)
11. ftp://www.3gpp.org/tsg\_ran/WG1\_RL1/TSGR1\_17/Docs/PDFs/R1-00-1428.pdf [↑](#footnote-ref-11)
12. http://pfister.ee.duke.edu/courses/ecen655/polar.pdf [↑](#footnote-ref-12)
13. [http://wireless.engineering.nyu.edu/static-homepage/tech-reports](http://wireless.engineering.nyu.edu/static-homepage/tech-reports/) [↑](#footnote-ref-13)
14. <https://www.o3bnetworks.com/technology/latency-throughput/> [↑](#footnote-ref-14)
15. With the satellite operating as a repeater rather than a relay (no on-board signal processing) [↑](#footnote-ref-15)
16. The Space Communication Protocol Standard <http://www.scps.org/> [↑](#footnote-ref-16)
17. Thales, Fraunhofer, Dish Networks and Ligado [↑](#footnote-ref-17)
18. https://portal.3gpp.org/desktopmodules/Specifications/SpecificationDetails.aspx?specificationId=3234 [↑](#footnote-ref-18)