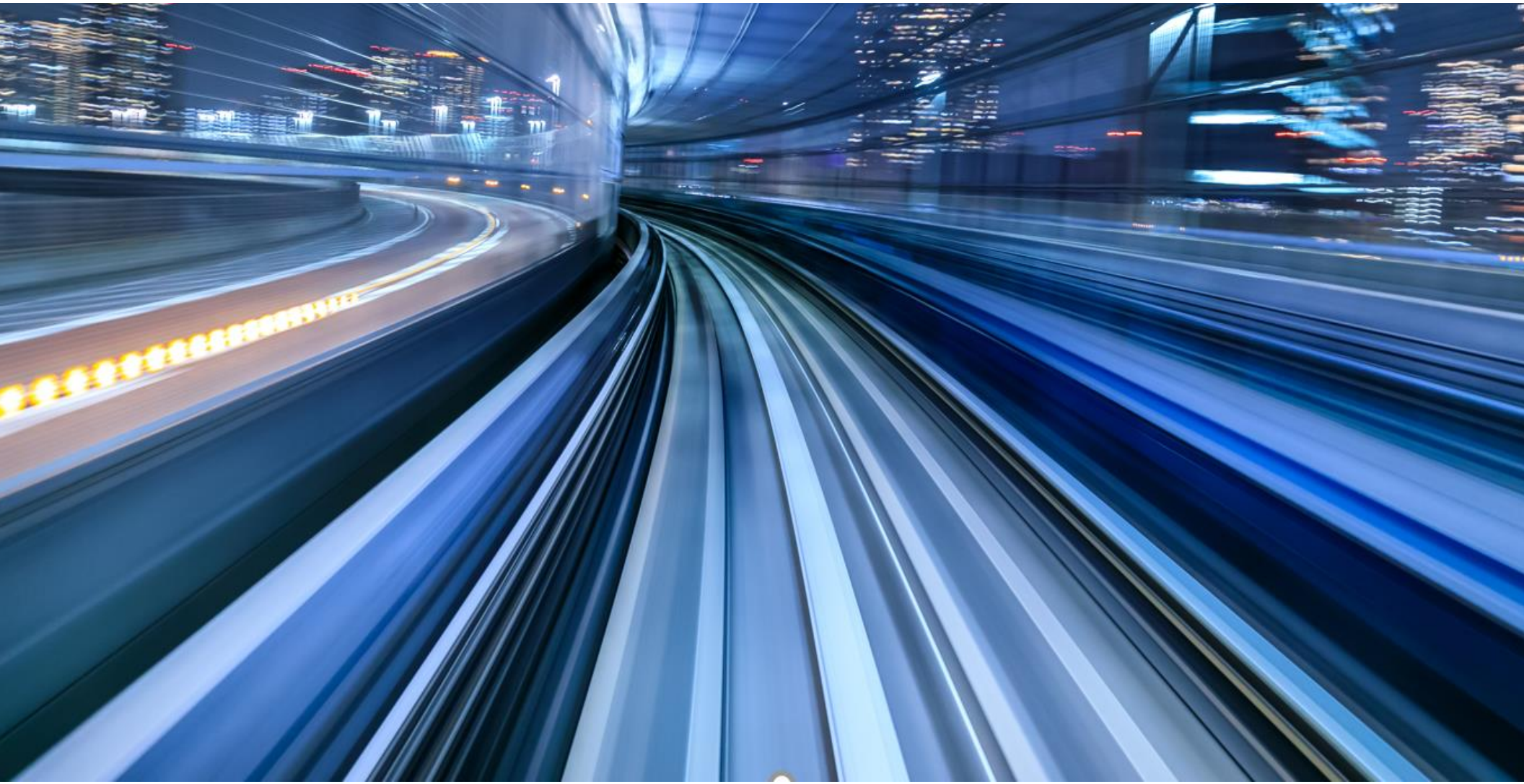
Latency matters

How to control latency and benefit from it

Monica Paolini, Senza Fili



1. An expanded role for latency

Latency, or delay, in wireline and wireless networks is nothing new, and, indeed it has shrunk as new technologies have emerged. Yet its impact continues to grow as networks become more complex, carry more traffic, and support more real-time services.

The role of latency in wireless networks is complex and multi-faceted.

On the user plane, it has a direct effect on subscriber experience, or QoE. High latency makes a voice conversation difficult, with users unable to understand each other or manage the conversation flow. It can have a crippling effect on video, because it degrades motion perception. Time-critical applications such as connected cars and AR simply do not work if latency is too high.

On the control plane, latency's impact on network performance is indirect but still causes a further erosion of QoE. High latency disrupts processing and transmission, and it reduces the efficiency in the use of network resources.

In 5G, extremely low latency in the RAN (on the order of 1 ms) is possibly the toughest requirement to meet. Some 5G use cases depend on low latency, and the increasingly complex and virtualized network architecture needs real-time traffic management to optimize network performance. At the same time, NR – New Radio, the 5G radio interface – edge computing, and network slicing will make it easier to reduce latency or manage it more efficiently.

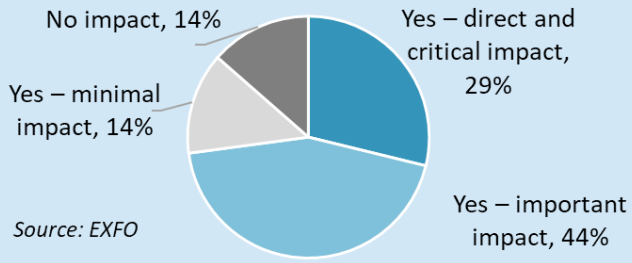
Ahead of 5G, the pressure for latency reduction is already on. Users expect the same QoE regardless of access type – wireline or wireless, licensed or unlicensed – and are quick to abandon apps that are too slow, or abandon the network and move to another operator.

Latency cannot be avoided, of course. But it can be precisely measured, understood, and managed in order to minimize its impact on QoE and network performance. Some aspects of latency are technology dependent and cannot be eliminated, some may be controlled or removed, and others are outside the network operators' control. In this paper, we will discuss how latency affects wireless networks and how operators can benefit from an efficient management of latency.

Latency matters for revenues

Service providers are aware that latency has a concrete effect on their bottom line. In a survey of service providers that EXFO conducted in 2017, 73% of respondents said latency has a critical and direct or an important impact on their revenues. They also included latency – or delay – among the three most important KPIs that determine QoE, along with dropped connections and throughput. Among respondents, 40% use a combination of one-way and two-way latency to measure latency, while 32% use two-way latency. At 42%, NTP is the most widely used technology for making these measurements, followed by GNSS/GPS (32%) and IEEE 1589 PTP (32%).

Does latency have an impact on your revenue stream?

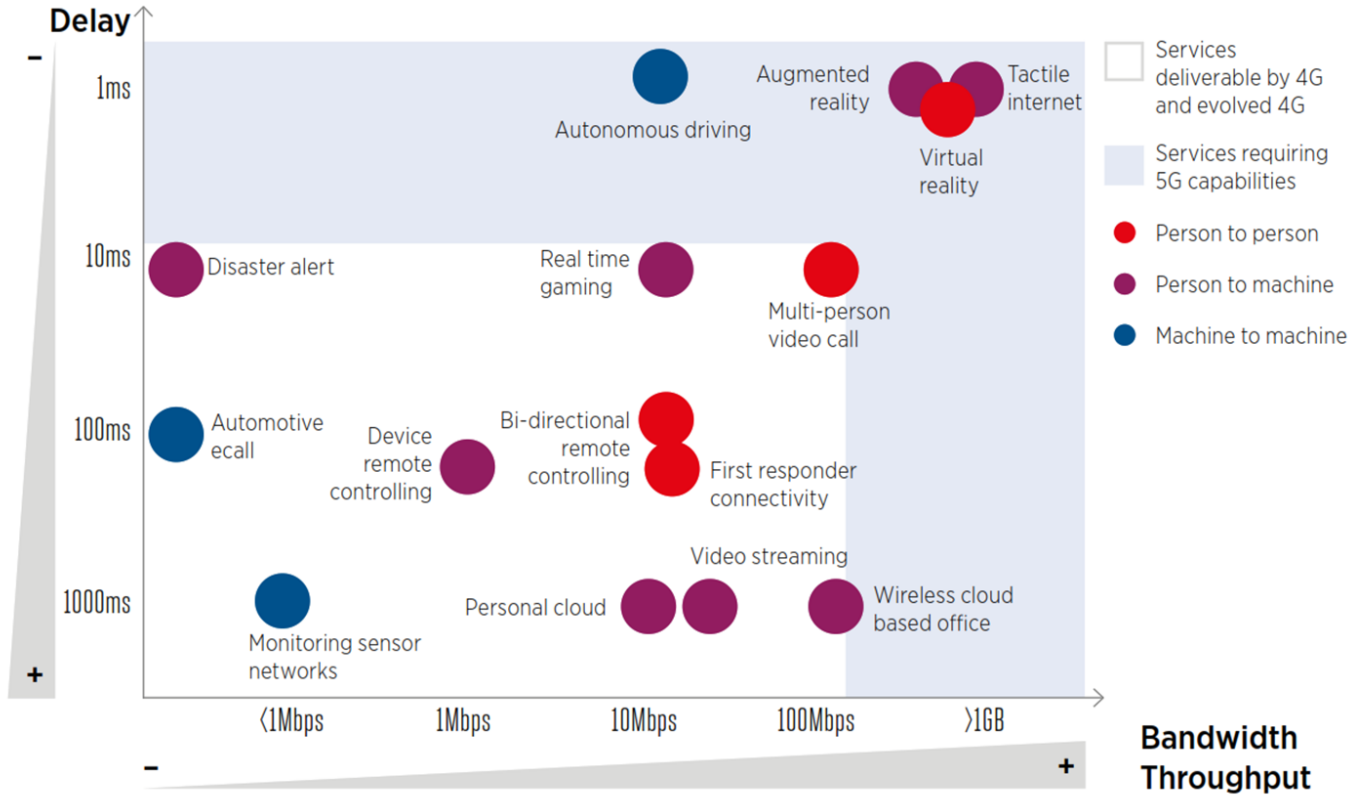


2. Where does latency matter?

Whatever subscribers do, latency matters. Nobody likes to wait. Nobody likes to get lost because the map app cannot load. Nobody likes to have the video of their kids freeze or get pixelated as they show it to their friends.

If all traffic were equally affected by latency, operators could only try to minimize latency across the network – an expensive effort likely to bring in only limited revenues. But the sensitivity to latency does vary greatly across applications and traffic type.

This gives operators the opportunity to differentiate how they manage latency across users and services, which opens the way to using low latency as a competitive, revenue-generating driver. In other words, operators can choose to reserve low-latency transmission for the applications that need it the most or that can generate higher revenues as latency decreases.



Source: GSMA Intelligence

Figure 1

The applications that are least sensitive to latency are those that do not have a real-time component. Many background tasks, such as email downloads, synching, and software updates, belong to this category, as do IoT applications that are based on sensor readings or monitoring.

Web browsing and applications such as social media, maps, and ride sharing are less severely affected by latency. There is evidence, however, that subscriber behavior is negatively affected when latency increases in purchasing and application use (see next section). It is difficult to quantify the impact, because often it cannot be directly linked to the activity on

the mobile device: an operator can track dropped calls or slow video from data from its network, but not missed purchases.

The applications where latency matters the most are those with a real-time component, where there is interaction among subscribers (e.g., a voice or video call, gaming), or timing is essential (e.g., autonomous driving, VR/AR, remote surgery). Applications such as video streaming, which account for a large part of wireless traffic, are not highly sensitive to latency because the impact of latency can be reduced or eliminated with techniques such as buffering.

Figure 1 shows the latency and throughput requirements of different applications or services. Use cases with tight latency and throughput requirements – such as autonomous driving, AR/VR, tactile internet, some healthcare applications, factory automation, and robotics – push us toward the top and right portion of the graph, where 5G is required. These are the applications that require ultra-reliable low-latency communications (uRLLC), with target latencies below 10 ms.

Luckily, only a limited selection of applications requires uRLLC, and operators can focus their latency reduction efforts on those. Extremely low latency can be expensive, especially if it has to be guaranteed to a large part of network traffic, so achieving it selectively is valuable.

The ability of mobile operators to support extremely low latencies where and when needed is crucial to the adoption of uRLLC services. Unless latency is sufficiently low, autonomous driving will not be safe for commercial deployments. VR/AR causes motion sickness if the latency is too high. Surgeons will refuse to operate remotely if the connection does not have low latency and low jitter. Enterprises will forgo investing in applications that do not meet their timing requirements.

"Latency has become one of the key characteristics that distinguish 5G from 4G or any other technology. The reduction of latency that's being made possible with 5G is quite extensive and variable.

"There is a range of different latency values that 5G makes possible. They allow us to broaden the services we offer into areas where we wouldn't have been capable before. It helps us expand into new markets."

Mansoor Hanif, Director of the Converged Network Research Lab, BT

3. Beyond wireless

Latency matters beyond wireless networks. Retailers, advertisers, and service providers closely monitor the impact of latency on user perception, behavior and, ultimately, purchasing decisions, because it can have a major impact.

Amazon found that a latency increase of 1 s cost it $1.6 billion in 2012 [4]. Today, the figure is likely to be much higher, because online sales have grown, while we are less patient with delays in connectivity.

Walmart found that delays of 1 s affect purchasing decisions. According to the Economist, in 2016, Walmart acquired Jet because of its "real-time pricing algorithm, which tempts customers with lower prices if they add more items to their basket. The algorithm also identifies which of Jet's vendors is closest to the consumer, helping to minimize shipping costs and allowing them to offer discounts. Walmart plans to integrate the software with its own"[12]. "It turns out that under a second was just too damned slow," notes journalist Thomas Friedman [5].

Sensitivity to latency in online gaming
>300 ms – game unplayable
>150 ms – player performance degraded
>100 ms – player performance affected
50 ms – target performance
13 ms – lower limit of detectibility

Source: PubNub

Table 1

Similarly, in 2009 Google found out that "increasing web search latency 100 to 400 ms reduces the daily number of searches per user by 0.2% to 0.6%" [1]. This is a small percentage, but it translated into 8 million canceled searches per day. Google also saw that "users do fewer searches the longer they are exposed." Latency not only changes the immediate behavior of subscribers, it conditions their long-term habits. If you know that video does not work well on your phone, you do not even open YouTube unless there is an urgent need to.

Microsoft found that a 1 s slowdown reduced queries by 1% and ad clicks by 1.5%. A 2 s slowdown reduced queries by 2.5% and ads by 4.4% [8].

Sensitivity to latency in games played over wireline networks also gives us some insight into the range of latencies we are sensitive to. According to PubNub [2] (Table 1), gamers are not able to detect latencies below 13 ms, and a 50 ms latency is sufficient for good QoE. While latencies above 50 ms are disruptive of subscribers' behavior, reducing latency to below this point may carry benefits that are noticeable only to skilled gamers.

This data shows how even small changes in latency can have a significant impact on the ability to keep users on a site, make purchasing decisions, or enjoy an application – and can translate directly into revenue gains and losses. Lowering latency, where possible, may make economic sense, but that cannot be known without estimating how much revenue the provider loses to high latency. Both figures are crucial to determining whether the new revenues make a good ROI case for investing in the necessary technologies.

The same holds true for mobile service providers, which need reliable ways to measure and monitor latency and subscriber behavior so they can assess latency's cost, invest in managing latency more efficiently, and choose solutions best suited to the specific response by their subscribers (e.g., the impact of latency on different subscriber segments or different applications).

4. Types of latency

Physics dictates that any signal transmission has an inherent latency. In wireless networks, there are multiple sources of latency, and they vary according to network conditions, traffic load, transmission problems, or network configuration. While it cannot be eliminated, latency can be minimized and managed.

Fixed latency can be estimated from the deployed network – e.g., topology, equipment, and technology (Figure 2). In the long term, fixed latency can be reduced through network upgrades and expansion – e.g., by moving to 5G, or deploying functions at the edge.

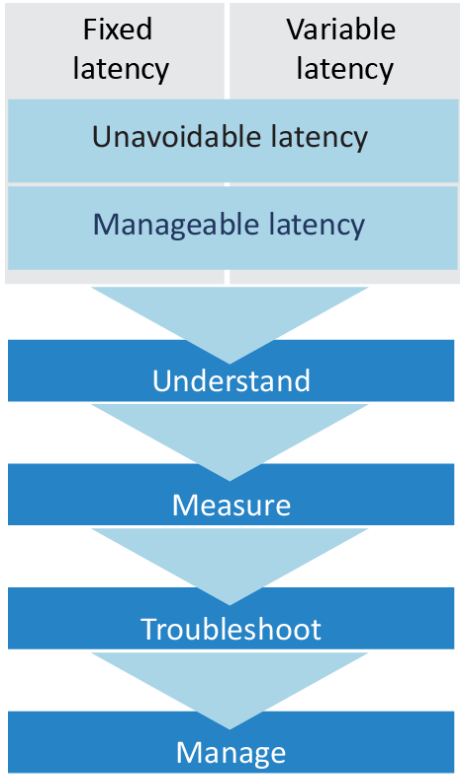
Unlike fixed latency, variable latency can be minimized and managed dynamically, to some extent. And unlike fixed latency, variable latency components have to be estimated in real time, because they fluctuate as users move around and as their usage patterns change.

Factors contributing to latency
Physical location and distance
Topology, architecture
Technology (e.g., air interface)
Transmission
Processing
Routing, switching
Traffic load, congestion

Table 2

Latency management is still in its early days, but automation, analytics, AI and, generally, more flexible and dynamic networks will increase the scope for latency reduction. To manage and minimize latency, it is crucial to identify factors contributing to it (Table 2) understand what it is, measure it, and monitor it. This information can then be used for troubleshooting and to locate and resolve any issue affecting performance. More generally, it can help operators maximize performance.

Fixed and variable latency



Source: Senza Fili

Figure 2

5. Sources of latency

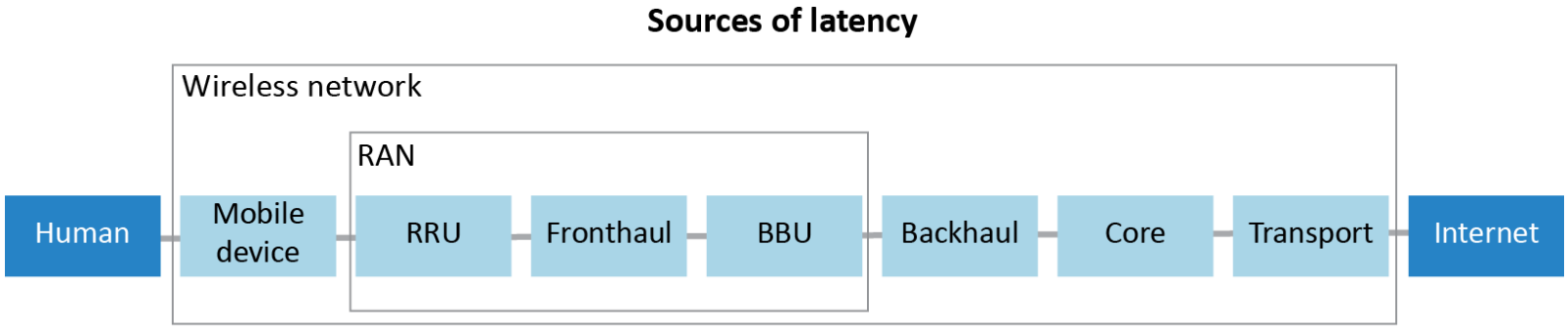
Usually, network latency is defined as the return latency or round-trip time (RTT) – the time it takes for a packet to be sent to the destination and for its receipt be acknowledged back at the source.

RAN latency is the component of end-to-end network latency that often attracts most of the attention, but there are multiple sources, internal and external to the wireless network that define network latency.

Figure 3 shows the sources that contribute to the end-to-end latency, from the time data gets to the mobile network to when it is delivered to a mobile device and received by a human. In an IoT private network, the two-way latency can be defined as the time it takes for some content or request to reach the IoT terminal and the receipt to be acknowledged by the local network.

Sources of latency include:

- The **subscriber** has to perceive, process and, possibly, act on the content delivered to the mobile device. Humans are slower than mobile networks, and this has a sizeable impact on perceptual and motor latency. It takes 10–15 ms for a visual input – e.g., a frame of a video in a mobile device – to reach the primary visual cortex, where the brain initially processes the information. It takes 40–50 ms for the brain to process motion perception, and more to reach the level of awareness needed to make a decision. The round-trip from visual stimulus hitting the retina to motor response (e.g., press a button) is more than 100 ms, and for most people around 150–200 ms.
- At the other end, there is a variable latency from the **internet** when retrieving data, or when a subscriber does a video call with another subscriber. Depending on the network condition and physical distance, the delay introduced by sources outside the network can be high. To reduce this latency, operators can cache content or move processing toward the edge of the network, but they have no control over the latency outside their network.



Source: Senza Fili

Figure 3

- Within the mobile network, starting on the subscriber end, the first component is the **mobile device**. For video content, display lag and frame rate contribute to the latency. For VoIP calls, encoding and decoding add to the end-to-end latency.
- Within the **RAN**, we saw a sharp reduction in latency as we moved from 2G to 4G (from 300–1000 ms in 2G to 10–100 ms in 4G). We expect another big drop, to 1 ms, as we move to 5G (Figure 4). In traditional distributed networks, where the **RRU** and **BBU** are at the cell site, the latency introduced by the backhaul is very low.
- As we move to C-RAN or, more generally, to virtualized RAN with remote BBUs, the **fronthaul** introduces a delay, which depends on the technology used (e.g., CPRI versus Ethernet) and the distance between the BBU and the RRU.
- The **backhaul** contribution to latency depends, like for the fronthaul, on the technology used, and on the distance between the BBU and

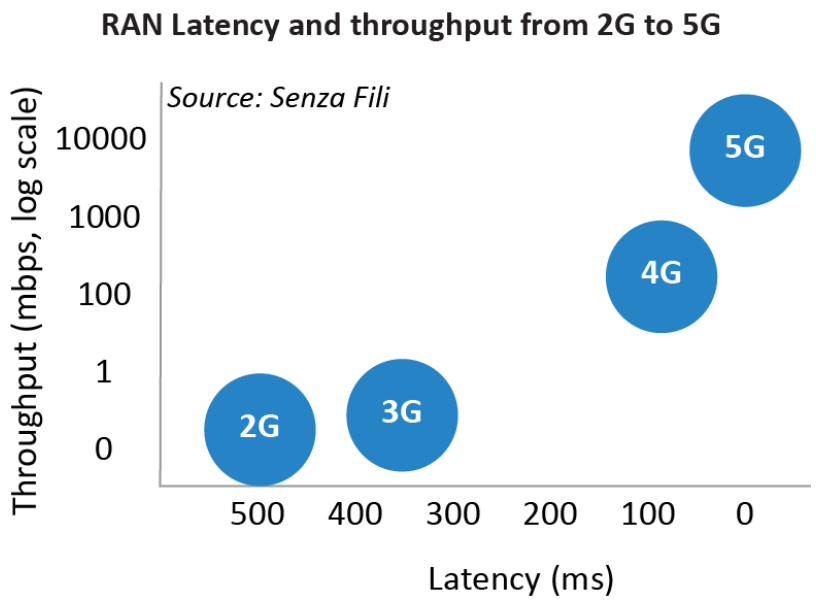
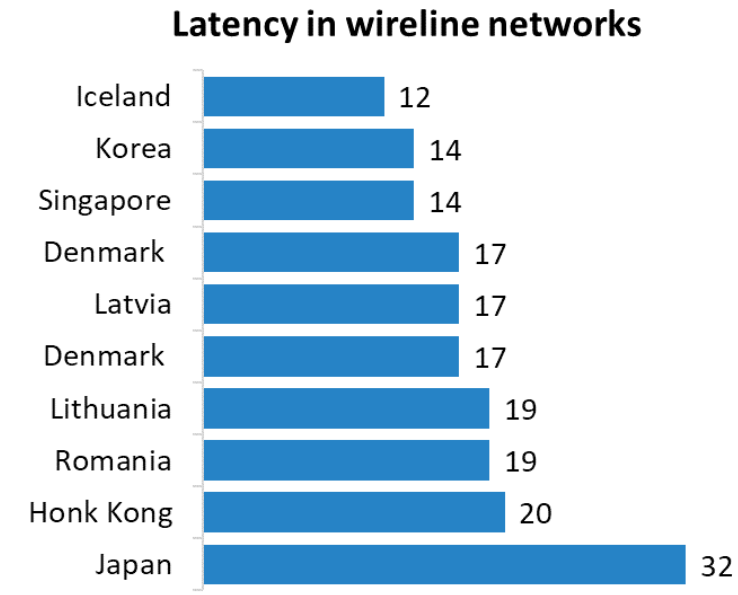
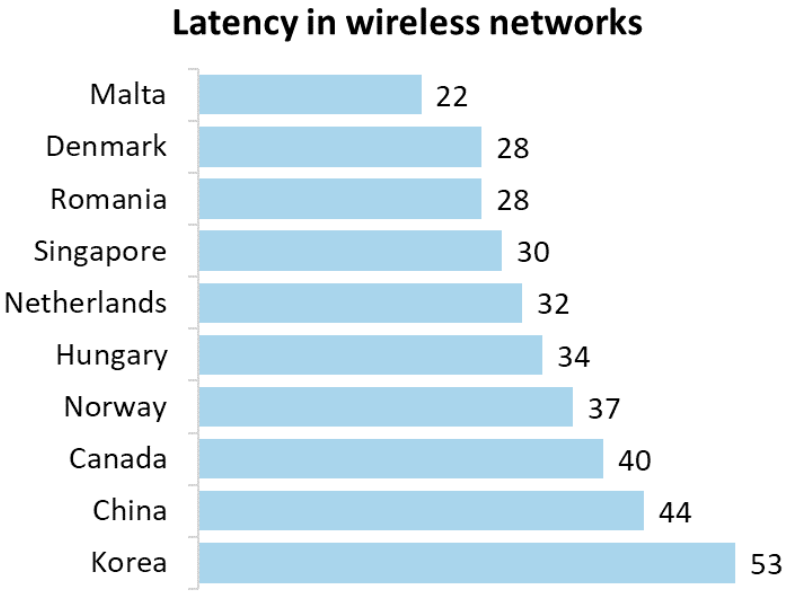


Figure 4



Source: Cisco Global Cloud Index, Ookla Speedtest.net/Ziff Davis



Source: Cisco Global Cloud Index, Ookla Speedtest.net/Ziff Davis

Figure 5

the core network (i.e., the EPC in LTE networks). Fiber adds 1 ms to the round-trip latency for each 100 km. This is tied to the speed of light and cannot be eliminated.

- The mobile **core** introduces a processing delay. As operators adopt virtualization and edge computing, this delay will depend on where content or an application is located physically. As a result, the contribution to latency from the core will depend on location as well as processing, and it will vary across functions and locations.
- **Transport** between the core and the internet and other networks adds a final latency component. This one depends on the link technology and distance.

The end-to-end round-trip latency varies across networks and on average across countries. Figure 5 from Cisco Global Cloud Index shows the latency for wireline and wireless networks, computed from the nearest web server available. Today, wireline latency is lower than wireless, but we expect that difference will shrink with the rollout of 5G.

6. Latency can cause more latency

High latency has a well-documented negative impact on QoE and perceived performance, as discussed above. It also has a secondary but insidious effect: it can further increase the perceived latency, especially in congested networks. It does this by triggering processes that are sensitive to delays and that, in turn, lead to an inefficient use of network resources.

TCP provides an example of the domino effect that latency and congestion can have. TCP is the protocol most commonly used for video streaming, because it can provide a reliable, high-quality video output. Given that, according to Cisco's VNI 2017, video accounted for 73% of IP traffic in 2016, TCP affects a huge portion of IP traffic [3].

TCP manages the quality of transmitted content by ensuring that transmitted packets are received and packet loss is avoided. The need to acknowledge packet receipt and to retransmit lost packets improves the consistency and reliability of the transmission, but creates additional processing and, hence, delays.

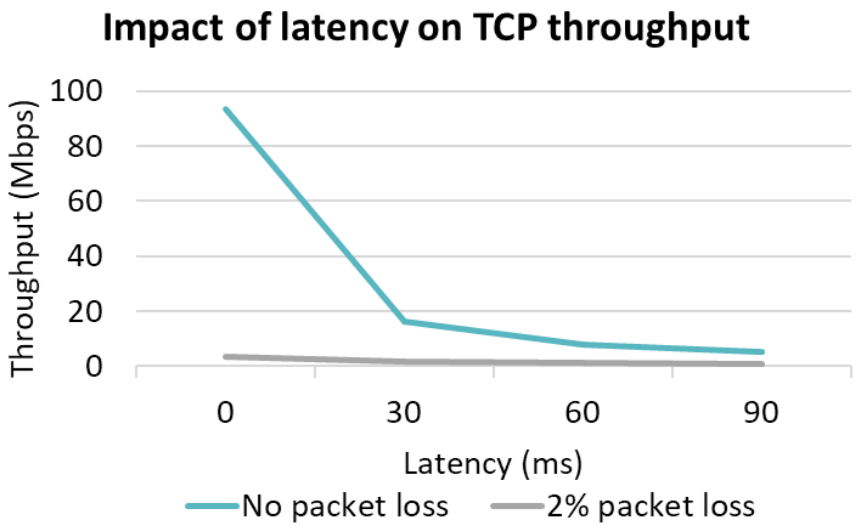
Before a new video frame is sent, TCP waits for the acknowledgment that the previous one has been received. This means that users experience video streaming at round-trip latency, even though content transmission is only on the downlink (DL).

The impact of the packet transmission process (Figure 6) on streamed video is exacerbated by the fact that the uplink (UL) is often slower than the DL. As a result, the perceived latency is higher than the DL latency multiplied by two.

When the traffic load is high, TCP messaging may cause congestion in the UL, and compromise or shut down the DL transmission, as well, as TCP waits for acknowledgments that are stuck in the UL. In mobile networks, this issue becomes even more severe, because video content is streamed, on average, in short segments which generate slower and more frequent TCP acknowledgments. Because the volume of TCP messages grows with traffic volume, TCP-driven congestion cannot be resolved only by adding the required capacity.

When TCP congestion occurs, subscribers perceive the network as slow or down, even though the air interface works fine. Their frustration may cause further disruption if they continue to resend video requests that further increase congestion in the UL. At the same time, for operators, it is not trivial to identify problems when they occur, because there is no obvious congestion in the DL. This becomes even more of an issue for video traffic which is predominantly DL but contributes to UL congestion through TCP messaging.

The TCP case illustrates the complex effect of latency on perceived performance and the need to understand and measure what causes it. This is crucial, because it enables operators to manage latency more efficiently – for instance, in this example, by optimizing TCP performance based on traffic conditions.



Source: http://smutz.us/techtips/NetworkLatency.html

Figure 6

7. Improving latency in wireless networks

Most wireless traffic comes from real-time applications and services that require low and reliable latency. As we move to 5G, there will also be growth in latency-sensitive traffic. uRLLC use cases will come of age, and they will increase the pressure to reduce latency across the network and, even more so, at the application or network-slice level. One way to do this is by managing latency. But before that, operators can reduce what we referred to as fixed latency by transitioning to 5G and by introducing new technologies that strengthen 5G or are enabled by it.

Table 3 lists 5G and emerging technologies that will lower latency, although for most of them the latency improvement is only one of the benefits. These new technologies will contribute to a greater efficiency in the allocation of network, which in turn will lead to a lower network latency, and a lower perceived latency.

For instance, lowering network latency increases the efficiency of TCP, and that triggers an increase in throughput and a reduction in perceived latency – along with an overall increase in QoE.

From the subscriber viewpoint, latency and throughput are tightly linked in their contribution to the user experience. Both high latency and limited throughput create the perception of a slow network, and, in most cases, the subscriber cannot identify the cause – latency, throughput, or both.

Operators are well aware of this. In the EXFO survey, throughput and latency are the second and third KPIs, after dropped calls, that affect QoE the most. With 5G, operators will have multiple ways to manage the complex interaction between bandwidth and latency.

Lowering latency in 5G networks			
RAN		**Core**	**Transport**
Frame/packet structure	RAN virtualization	SDN	Fronthaul and backhaul evolution
Waveform, multiple access	mmW for access	NFV	Wireless-wireline convergence
Modulation and coding	Location-aware communication	Edge computing (MEC, Fog, caching)	mmW for backhaul
Transmission	Reinforcement of QoS and QoE		
Control channel	Separation of control and user plane		
Symbol detection	TCP optimization		

Source: Parvez et al., Senza Fili

Table 3

8. From two-way to one-way latency measurements

RTT is the most common way to measure latency because it is easy to calculate: it is computed at a single location, as the difference of two timestamps generated by the same equipment.

While the RRT is a useful indication of overall latency in a network and can be used to give a measure of jitter (variability in latency), it is not sufficient for capturing both the latency that users experience, and the impact latency has on network performance.

The reason is simple. Networks are asymmetric (DL and UL have very different characteristics, and latency is among the ones that differ), and our use of networks is asymmetric. There is more DL traffic than UL traffic, and the perceptual impact of latency in received versus sent traffic may be different.

Network asymmetry would not be an issue if the relation between UL and DL latency were constant and could be predicted based on network topology and architecture. However, this is not the case. Latency fluctuates due to multiple factors and the interaction among them, and they do not all have the same impact on UL and DL.

To measure latency accurately, it is imperative to measure the one-way latencies for DL and UL separately, and to be able to troubleshoot and manage them independently. For instance, a network may have acceptable latency, but the QoE on streamed video may be low because the network has a relatively high DL latency. In other cases, a slow UL may create congestion in both UL and DL if TCP messages overload the UL channel.

"The monitoring and measurement tools have to be very precise in measuring the latency each way. We cannot typically calculate the one-way latency simply by dividing the RTT, the round-trip time, by two. Latency needs to be calculated in each direction to identify the precise bottlenecks and address them."

Neeraj Pandey, Associate Vice President, Vodafone Shared Services, Vodafone

Yet most operators do not measure one-way latency. According to EXFO's survey, only 45% of operators use one-way latency to monitor and track service performance. This is understandable. Measuring one-way latency with the reliability and the microsecond accuracy that operators need has traditionally been difficult and expensive: it required complex solutions such as NTP, GNSS/GPS and IEEE 1589 PTP, which are based on external clock synchronization. Also, external timing sources with solutions like GPS do not work reliably, or at all, in indoor locations or where there is no line of sight to the timing source.

A simpler, less expensive way to measure one-way latency is to compare timestamps from equipment elements in the wireless network, without relying on external clock synchronization. Such measurements are more more reliably available across the entire footprint, including locations that are indoor or not in line-of-sight with an external synchronization source, provided that timestamps from the equipment elements involved can be synchronized.

9. Measuring latency at the application or service level

Moving beyond RTT to measure one-way latency is crucial, but it is only the first step toward understanding and managing latency.

The next step is to measure latency at the application or traffic-type level. Latency requirements vary across applications and types of traffic, and operators can use this to their advantage. Users are more tolerant of high latency and jitter when they browse the web than when they watch a video on Netflix. With voice or video calls, users are highly sensitive to latency and jitter.

Until recently, latency measurement at the application level has been helpful for understanding QoE, but not critical. Operators could only try to lower latency for the entire channel, so a decrease in latency would equally benefit all types of traffic. Increasingly, though, network operators can manage traffic at the application, service, and traffic-type levels with QoS, network slicing, and edge computing, or by directing different traffic to best-suited access networks (e.g., LTE versus Wi-Fi).

This allows operators to fine-tune the allocation of network resources. They can minimize latency for applications that require low latency, and balance that by loosening the latency requirements for other applications. To do this, operators need to be able to measure one-way latencies accurately and reliably for each relevant traffic class or network slice, both to decide how to manage traffic and to measure the outcome.

Accurate one-way measurements at the application, service or slice level are also crucial for edge computing and function virtualization. Once the operator knows the amount of latency and its sources for each traffic flow, it can decide whether to move the processing associated with that traffic flow to a new location in the wireless network – typically, toward the edge. This puts content storage and processing closer to the device, and it has a predictable, positive effect on latency and jitter.

Because some components of latency vary with traffic load and network conditions, constant monitoring of latency will help operators refine network use of resources in real time or near-real time. For instance, during a sports event, an operator may decide to target all its latency-reduction efforts to the sports venue and surrounding areas, reallocating virtualized hardware resources for the duration of the event.

"What we'll be looking for is a way to provision various checkpoints across many parts of the network, where our customers will be. Then to measure the aggregate latency to those end points and be able to define a delta at which we expect all our latency to be managed. If we're falling out of that latency, the measurement tool needs to send us alerts. If we're getting close to the edge, send us alerts, which would then need to trigger the network slicing algorithm, in certain cases, to allocate more resource to bring that latency down for that particular customer."

Mansoor Hanif, Director of the Converged Network Research Lab, BT

"You have to take into account what kind of latency should be there for the control plane, for the user plane, for the synchronization plane, and for the management plane. All these latencies need to be defined and measured across the network. And before benchmarking the network, you should be very sure in what areas you can minimize the latency."

Neeraj Pandey, Associate Vice President, Vodafone Shared Services, Vodafone

10. Managing latency

High latency carries a hefty cost. On the subscriber side, high latency creates a bad user experience, frustration and even long-term reduction of service use. The effects are well known: low QoE scores, lost revenues, increased churn. On the network side, high latency leads to an inefficient use of network resources: as latency grows, network performance goes down, and the operator extracts less value – in terms of money spent per bits transmitted – from the existing infrastructure. In other words, high latency lowers the ROI.

Reducing latency is good, but it is also expensive. Moving, to the lowest end-to-end network latency that technology allows is prohibitively expensive. It makes for impressive demos, but unsustainable commercial deployments. An operator would have a hard time justifying the required investment against the revenues it may generate.

Luckily this is not required. Only a few use cases require low or extremely low latency. Some of those do not require much bandwidth, so their contribution to the traffic load is minimal. Others, such as VR/AR, have high bandwidth requirements but may also be hosted at the edge, reducing the challenges and cost of latency reduction.

As operators move to 5G, lowering latency becomes less challenging and more affordable if they lower latency only to the extent that is financially sustainable, and manage latency aggressively only for the use cases that require it.

This will create a differentiated latency environment in which the operator has the flexibility to assign specific latency levels to different traffic flows, network slices, applications, or users. For instance, an operator may create a service plan for gamers in which the subscriber pays more for a guaranteed lower latency than a regular subscriber who receives best-efforts latency. For enterprise applications and IoT, an operator may gain additional revenues by offering SLAs that can support very low latency requirements. The ability to manage latency can be a valuable asset that enables the creation of new services, generates new revenues, and differentiates an operator's offering from the competition's.

"Reliability is key. It's not so much trying to get the latency as low as possible; it's about being able to manage the latency at a specific, reliable level, and being able to maintain that latency target for a specific user. That's why we talk about ultra-reliable low-latency communications: because we need to be able to specify latency for a certain customer, and then guarantee that latency over a specific period."

Mansoor Hanif, Director of the Converged Network Research Lab, BT

Even more importantly, however, operators have to move beyond RTT latency. 5G gives operators access to an increasing array of tools to understand, measure, troubleshoot and manage latency that they can use to improve network efficiency and QoE, and create a new class of services. To take advantage of these tools and to extract revenues from latency reduction, operators have to switch to one-way latency for KPIs and other network metrics – for the entire channel and separate traffic flows (defined by application, service, network slice, or subscriber), and in near-real time or real time. Moving to one-way latency is the first step that operators have to take to manage latency and to reap the performance and financial benefits of the lower latency that 5G promises.

11. Implications

Latency is gaining a prominent role in defining and improving QoE, in protecting and generating revenues, and in improving network performance and efficiency.

High latency has a negative impact on subscriber experience. Not only do voice calls and video quality degrade, but latency affects subscriber behavior (e.g., app utilization or purchasing decisions), even for non real-time traffic.

With 5G, latency will become even more central to improving QoE and to supporting uRLLC and other use cases.

While latency affects all traffic and QoE, not all applications or use cases require or benefit from extremely low latency.

Operators do not need to operate their entire networks at extremely low latencies. They can selectively manage the latency at the application, network slice, traffic flow, or user level. This results in a cost-effective way to allocate network resources, improve QoE and generate new revenues.

To manage latency, operators need to understand what the sources of latency are, measure them, and use the data to troubleshoot and optimize their networks. But to do so, they need to move beyond round-trip latency. They have to track one-way latency separately in the downlink and in the uplink, and do so for different traffic flows.

Glossary

2G	Second generation
4G	Fourth generation
5G	Fifth generation
AR	Augmented reality
BBU	Baseband unit
CPRI	Common public radio interface
C-RAN	Cloud RAN
DL	Downlink
eNB	Evolved NodeB
EPC	Evolved Packet Core
GNSS	Global navigation satellite system
GPS	Global positioning system
IEEE	Institute of Electrical and Electronics Engineers
IoT	Internet of things
IP	Internet Protocol
KPI	Key performance indicator
LTE	Long Term Evolution
mmW	Millimeter wave
NFV	Network Functions Virtualization
NR	New Radio
NTP	Network Time Protocol
PTP	Precision Time Protocol
QoE	Quality of experience
QoS	Quality of service
RAN	Radio access network
ROI	Return on investment
RRU	Remote radio unit
RTT	Round-trip time
SDN	Software-defined networking
SLA	Service level agreement
TCP	Transmission Control Protocol
UL	Uplink
uRLLC	Ultra-reliable low-latency communications
VoIP	Voice over IP
VR	Virtual reality

References

[1] Brutlag, Jake, Speed matters for Google web search, Google, 2009.
[2] Burger, Thomas, How fast is realtime? Human perception and technology, PubNub, 2015.
[3] Cisco, Cisco Global cloud index supplement: Cloud readiness regional details, 2017.
[4] Eaton, Kit, How one second could cost Amazon $1.6 billion in sales, Fast Company, 2012.
[5] Friedman, Thomas L., Thank you for being late: An optimist's guide to thriving in the age of accelerations, Farrar, Straus and Giroux, 2016.
[6] GSMA, Unlocking commercial opportunities: From 4G evolution to 5G, 2016.
[7] International Telecommunication Union, IMT Vision – Framework and overall objectives of the future development of IMT for 2020 and beyond, Recommendation ITU-R M.2083-0 (09/15).
[8] Noction, Understanding the impact of network latency on service providers' business, 2012.
[9] Paolini, Monica, Mastering analytics: How to benefit from big data and network complexity, 2017.
[10] Paolini, Monica, Power at the edge: Processing and storage move from the central core to the network edge, Senza Fili, 2017.
[11] Parvez, Imtiaz, A survey on low latency towards 5G: RAN, core network and caching solutions, IEEE Access, 2017.
[12] The Economist, Boxed-in unicorn, 2016.
[13] Thierno Diallo, Deliver superior quality of experience through accurate one-way delay assurance, EXFO, 2017.

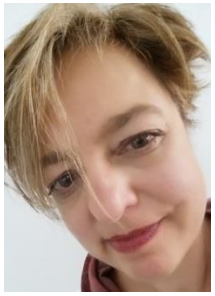
About EXFO

EXFO develops smarter network test, monitoring and analytics solutions for the world's leading communications service providers, network equipment manufacturers and webscale companies. Since 1985, we've worked side by side with our customers in the lab, field, data center, boardroom and beyond to pioneer essential technology and methods for each phase of the network lifecycle. Our portfolio of test orchestration and real-time 3D analytics solutions turn complex into simple and deliver business-critical insights from the network, service, and subscriber dimensions. Most importantly, we help our customers flourish in a rapidly transforming industry where "good enough" testing, monitoring and analytics just aren't good enough anymore—they never were for us, anyway. For more information, visit EXFO.com and follow us on the EXFO Blog.

About Senza Fili

Senza Fili provides advisory support on wireless technologies and services. At Senza Fili we have in-depth expertise in financial modeling, market forecasts and research, strategy, business plan support, and due diligence. Our client base is international and spans the entire value chain: clients include wireline, fixed wireless, and mobile operators, enterprises and other vertical players, vendors, system integrators, investors, regulators, and industry associations. We provide a bridge between technologies and services, helping our clients assess established and emerging technologies, use these technologies to support new or existing services, and build solid, profitable business models. Independent advice, a strong quantitative orientation, and an international perspective are the hallmarks of our work. For additional information, visit www.senzafiliconsulting.com, or contact us at info@senzafiliconsulting.com.

About Monica Paolini



Monica Paolini, PhD, founded Senza Fili in 2003. She is an expert in wireless technologies, and has helped clients worldwide to understand technology and customer requirements, evaluate business plan opportunities, market their services and products, and estimate the market size and revenue opportunity of new and established wireless technologies. She frequently gives presentations at conferences, and she has written many reports and articles on wireless technologies and services. She has a PhD in cognitive science from the University of California, San Diego (US), an MBA from the University of Oxford (UK), and a BA/MA in philosophy from the University of Bologna (Italy). You can contact Monica at monica.paolini@senzafiliconsulting.com.