IEEE P802.24

Vertical Applications Technical Advisory Group

Project	IEEE P802.24 Vertical Applications Technical Advisory Group
Title	**Low Latency Communication White Paper**
Date Submitted	2023-07-13
Source	Oliver Holland Allan Jones Tim Godfrey Ben Rolfe Dillon Seo
Re:	N/A
Abstract	This contribution provides a first version of the Table of Contents of the Low Latency Communication White Paper. It will be updated (along with this Abstract) as the content materializes and is included.
Purpose	Assist in the development of the Low Latency Communication White Paper
Notice	This document has been prepared to assist the IEEE P802.24. It is offered as a basis for discussion and is not binding on the contributing individual(s) or organization(s). The material in this document is subject to change in form and content after further study. The contributor(s) reserve(s) the right to add, amend or withdraw material contained herein.
Release	The contributor acknowledges and accepts that this contribution becomes the property of IEEE and may be made publicly available by P802.24.

1. Background and Introduction

This white paper is to inform users and IEEE 802 working groups on the applications and requirements for low latency communications. Low latency is challenging to implement in wired or wireless networks that communicate over a shared medium. Wireless networks that operate in unlicensed spectrum with contention-based protocols make low latency more difficult to achieve.

Low latency is typically achieved by a combination of access control and scheduling along with increasing bandwidth (overprovisioning) in the network.

2. Low Latency Communications Applications

The need for low latency communication is being driven by a group of application requirements. A set of representative applications are described below, but new applications with low latency requirements continue to emerge.

Electric Utilities - Grid Protection

The utility is considered an entity (or entities) that manage the distribution of electricity on the transmission grid and the distribution grid. The power distribution network involves substations, and various protective and control devices that communicate over communications networks.

Low latency or “real-time” performance of the network is important for specific grid use cases and applications.

Ethernet (carried over fiber and copper) is widely used for this application. The real-time behavior of Ethernet based communication networks is defined in IEC 61784-2. There are 6 (plus one technology specific) consistent sets of parameters described to define the requested and achieved Real-time Ethernet behavior of end-to-end stations. For the network components, using TSN is an effort ongoing in IEC SC 65C.PT61784-6, dealing with a TSN profile for industrial automation applications. The application of IEEE 802.1 TSN for utilities is the topic of a prior white paper [1].

A leading grid application for low latency is protection. Protective relays protect electrical transmission lines against fault conditions (line down, short circuits between conductors or to ground). Simple protection schemes measure voltage and current at one end of the transmission line. Differential protection schemes determine fault conditions by measuring real-time differences in voltage and current between the ends of the line. This requires an independent communication link with very low (<10mS) end to end latency to carry the measurements between the relays at the ends of the line. The communication link latency must be highly consistent and predictable. The latency requirement is less than one cycle of the AC waveform (16.6 mS, or 20 mS), because time must be allowed for the mechanical operation of the relay in the case of a fault.

The communication link connection is typically fiber, although copper circuits are also used. Power Line Carrier and point to point microwave are less commonly used.

While the highest voltage transmission lines are likely to rely on fiber due to its reliability and predictability, there are other less critical protection applications where low latency wireless can offer a solution.

Direct Transfer Trip (DTT) is a protection scheme often used to connect medium to large scale Distributed Energy Resources (DER) systems (such as wind farms and solar arrays) into the distribution grid (between 4 and 35 kV). Low latency is required because the fault detection system sends commands to remote breakers. A delay in the “disconnect” command can cause damage due to the fault current. DTT is also used for “anti-islanding” protection, to disconnect a DER system from the main distribution feeder if the main feeder has an outage. This prevents “backfeeding” electricity into a feeder that should not be energized from the DER system.

A third application for low latency is wildfire protection. In areas that are susceptible to wildfires, there is a risk from energized conductors falling to the ground and starting a fire because of wind or other events. Low latency communications from sensors to circuit breakers can be used to identify a break or fault, and de-energize the circuit before the conductor hits the ground.

Low-latency Security Requirements

Low latency for networks in regard to security becomes even more important; especially due to recent changes in how people work remotely and emerging technologies.

Securus Communications[6] points out 5 reasons why low latency is important for today's networks.

1. Nextgen Voice and Video Services have created a unprecedented low-latency demand on current networks. High Definition 4K/8K streaming accommodating remote work requires high bandwidth and low latency to make these experiences as seamless as possible. Providing secure communications on top of the base requirements puts an even greater strain on low latency requirements.

2. Real-Time Retail Customer Analytics, is another reason low-latency networks are required. Companies try to identify customer trends in real-time. This requires low-latency networks. A combination of AI algorithms and real-time analysis often happening before the customer leaves the store after checking out with their purchase is pushing low-latency and security requirements beyond previous levels.

3. Industrial Internet of Things (IIOT) where secure communications between massive scale devices providing analytics and control on a level never seen is pushing low latency in critical control systems.

4. Autonomous vehicles have also been pushing Multi-access Edge Computing (MEC) which is only enabled by low-latency networks. Secure communications are critical for this function as human safety is involved and real-time analysis of vehicular traffic is critical in this role.

5. Virtual Reality and the Metaverse is one of the latest emerging technologies that requires real-time secure communications as people use AR/VR headsets to intercommunicate across virtual worlds. Low-latency and security is essential in providing a smooth unincumbered experience for the potentially massive users interacting with each other across large geographic distances.

In addition to the above highlighted use cases involving secure low latency communications, there is another often overlooked area involving Medical IoT devices. A paper published by the IEEE[7] points out these issues. The paper points out that within the scope of healthcare applications, delay would form a dangerous risk in case the system does not meet the compatibility requirements of health monitoring, in addition to the several security and privacy threats that are encountered. To ensure the safe transmission of data between IoT devices and the cloud, while keeping the possible network latency and response time to a minimum, the present study proposes a three-layered IoT-Fog computing model that deploys an authentication stage and an encryption stage with cloud computing.

Given the above use cases, its clear that we can't just look a low-latency through a single lens and that current use cases require us to look at secure low-latency solutions.

Real-time Mobile Gaming

Real-time mobile gaming is a fast-developing application category. Different from traditional games, real time mobile gaming is very sensitive to network latency and stability.

The mobile game can connect multiple players together in a single game session and exchange data messages between game server and connected players. Real-time means the feedback should present on screen as users operate in game. For good game experience, the end to end latency plus game servers processing time should not be noticed by users as they play the game.

The challenges that real-time mobile gaming encounter is the worst-case latency. Since the high latency spike is highly likely to cause packet loss and packet disorder, hence impact quality of experience. [4]

Wireless Console Gaming[4]

Console gaming involves various genres of games, but the main genre we are focusing on is latency sensitive online FPS (First Person Shooter) games. This is an interactive gaming experience with real-time feedback and response. A Synchronized game state is established among players in the same match to get the best performance. FPS gaming is centered around guns and other weapon combats in the first-person point of view with which the player sees the action through the eyes of the player character.

In multiplayer FPS game, more than one person can play in the same game environment at the same time either locally or over the internet. Multiplayer games allow players interact with other individuals in partnership, competition or rivalry, providing them with social communication absent from single-player games. In multiplayer games, players may compete against two or more human contestants, work cooperatively with a human partner to achieve a common goal, supervise other players' activity, co-op. Multiplayer games typically require players to share the resources of a single game system or use networking technology to play together over a greater distance.

Playing online on a console has 2 types of internet connectivity, which is either wired or Wi-Fi. Most of the gaming consoles today support Wi-Fi 5. But Wi-Fi has an especially bad reputation among the gaming community. The main reasons are high latency, lag spikes and jitter. According to a top-selling online console game in the US up to 79% of FPS players are using Wi-Fi connected consoles. [4]

Cloud Gaming

Cloud gaming is another type of video game potentially played on light-weight devices at users premise. Unlike other gaming hardware, user devices do not need to render pictures or video. Instead, they are rendered at the cloud server. The picture/video generated at the cloud server are streamed to the user devices, and the user devices just display the received picture/video on its display. The cloud game can accommodate and connect multiple players in a single game session just as mobile gaming scenario.

The cloud gaming requires low latency capability as the user commands in a game session need to be sent back to the cloud server, the cloud server would update game context depending on the received commands, and the cloud server would render the picture/video to be displayed at user devices and stream the picture/video content to the user devices. This cycle needs to be short enough so users do not feel lagging responses.

With cloud gaming experience, users can play large amount of game titles as they will be provided and hosted by the cloud server. Users can pick up game title from the library on the cloud server. Another benefit of the cloud gaming is that the user device could be light-weight in terms of hardware footprint. The user devices only need to decode and display received picture/video content. This way, users can enjoy realistic and immersive game experience without requiring heavy computation at user devices. The light-weight user device leads to lower cost and longer battery life, which could motivate gamers to play on the games more. [4]

Industrial Systems

Industrial systems include a wide range of applications: process monitoring, automation, control systems, human-machine-interfaces (HMI), Automated Guided Vehicles (AGVs), robotics and AR/VR. Recently, several standard developing organizations have published detailed description of industrial application and their requirements, such as:

- **IEEE 802.1 NENDICA Report Wired/Wireless Use Cases and Communication Requirements for Flexible Factories IoT Bridged Network (**802.1-18-0025-06-ICne);
- **IEC/IEEE 60802 Use Cases for Industrial Automation (TSN-IA Profile for Industrial Automation);**
- **3GPP TR 22.804 Technical Specification Group Services and System Aspects; Study on Communication for Automation in Vertical Domains.**

The purpose of this document is not to repeat the detailed application descriptions, which can be found in above references. Instead, the focus is to summarize the challenges and requirements of real-time and time-sensitive applications that are most relevant to IEEE 802.

Many industrial applications can be considered delay-tolerant (e.g. process monitoring, industrial sensor networks, etc.) with latency requirements in the order of 100msec or more. Such applications may be served by existing wireless standards and are not considered in this report. This report focuses only on time-sensitive and real-time applications. [4]

Real-time video

Today, many devices handle video streaming via 802.11 wireless LAN. Most of them are not latency sensitive. However, some video applications require low latency capability, when the application provides interactive play. Example of such applications includes AR/VR, and video cable replacement [3].

In many of these cases, the latency requirements are derived from the video frame rate. As of today, 60Hz framerate is commonly used, i.e., 16.7msec per frame. However, it is possible that the video rendering system would migrate to high frame rate solution, i.e., 120Hz which resulting in 8.33 msec per frame, etc., in the future.

To accommodate end-end signal processing in a video frame, the signal processing delay plus transmission latency need to be less than 16.7 msec. For these applications, ideally, 10[msec] one-way or roundtrip delay should be considered as a targeted specification for the radio link transmission, allowing 6.7msec for other signal processing including, but not limited to, video signal encoding (compression), in-device frame forwarding, video signal decoding (decompression), etc.

When the video frame rate of 120 Hz (8.33msec per frame) is used, ideally, 3 msec delay should be considered as a target for the radio link transmission, allowing 5.33 msec for other signal processing.

The following figure depicts the difference between a video application which does not require low latency capability and a video application which requires low latency capability. In general, low latency requirements arise when there is a control loop in the system. [4]

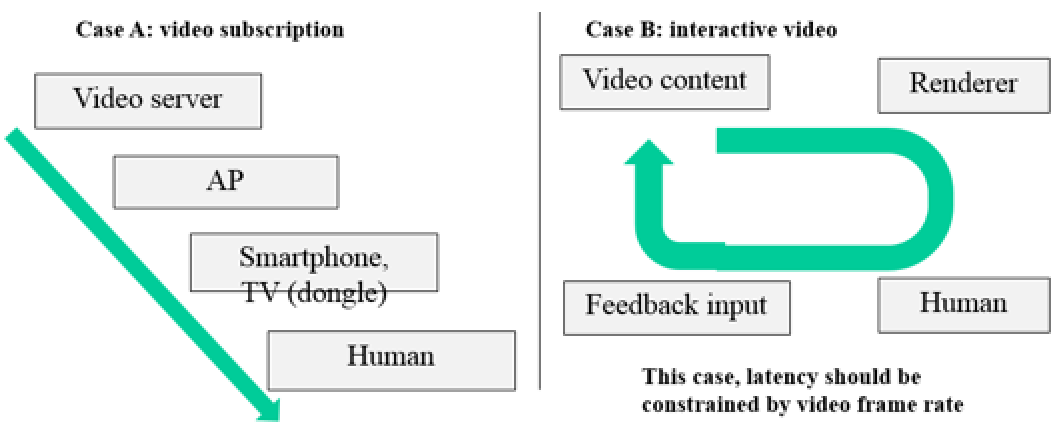

Figure 2-1 Difference between buffered video and live video

Drone Control

Drone is an aircraft without a human pilot aboard. Drones are rapidly popularized and utilized for a wide array of uses. Gartner mentions that worldwide production of drones neared 3 million units in 2017 [8]. Wi-Fi has an important role to control drones by providing following functions.

- **Tele control**

Controlling motions and functions of the drone. A few Kbps of data rate is required.

- **Data transmission**

Monitoring information from sensors in a drone or information of the status of the drone itself. A few Kbps~Mbps of data rate is required.

- **Picture / video transfer**

Transferring recorded pictures or videos by the drone. More than tens of Mbps of data rate is required. [4]

AR/VR

Use Cases: There are a number of AR/VR use cases that are expanded upon in the 802.21 document “Network Enablers for seamless HMD based VR Content Service” [5]. We won’t replicate these here in this whitepaper, but we can refer to the appropriate document found in the reference section.

Network Requirements

The network requirements for AR/VR can be summarized in the table below. For more detail the report on AR/VR Use Cases and Enablers can be found in the reference section. [5]

		VR HMD Requirements
Data transmission rate		~ 20 Gbps
Latency		~ 5 ms (at wireless medium), 20 ms (motion-to-photon/audio)
Jitter		< 5 ms
Transmission range	Indoor	5 m
	Outdoor	Several hundred meters
Mobility	Indoor	Pedestrian speed < 4 km/h
	Outdoor	200 km/h
PER		10^{-6}

Table 2-1 - VR Requirements

3. Performance Requirements for Low Latency Communication

Derived from the discussion on applications in Section 2 and also using other sources such as the ITU definition of URLLC, will list the performance requirements of low latency communication such as:

- End-to-end data transfer latency (Edge to Edge)
- Session establishment latency(?)

- Perhaps radio access latency (noting that in some fora, this distinction is made) E.G. use cases with edge intelligence where the device to edge computing service is the critical path.
- Reliability, noting that many applications also have this requirement
- Data capacity (identify trade-offs between achieving low latency and most efficient use of bandwidth)
- Synchronization among flows (e.g., with audio/video for haptic+AV applications…?)
- What is the opportunity for networks to retry lost packets? How does this vary for different applications and use cases?
- Describe the relationship between reliability requirements and data rate. Not all low latency applications require high bandwidth, but the application demands very high reliability (in terms of meeting the latency requirement)
- Some applications have a requirement for precision in the haptic feedback (precision is related to low latency – delay results in error)

4. Key Technologies/Solutions Supporting Low Latency Communication

Summarizing those technologies that have to be considered/utilized in order to achieve low latency, often in conjunction with high reliability. For example:

- Changes to framing to minimize wait time to receive a frame before processing the frame
- Rendering of video can be optimized based on the importance of the image, and whether the user’s eye is looking in that direction. This can allow lower latency overall.
- Video interpolation can potentially compensate for bandwidth limits that would otherwise limit frame rate.
- Prioritization of data within an application can ensure that the most user-perceptible aspects are provided the lowest latency handling in the overall system.
- Softwarization to optimize communication path through invoking elements in software at better locations?
- Network sharing to optimize communication path; neutral hosting, etc., etc.
- Multi-connectivity (as a means to still achieve reliability while reducing latency—noting that many low latency applications also require a vast *increase* in reliability compared with what is currently achieved (at least wirelessly))
- New coding approaches to achieve latency and high reliability
- New protocols

- Others (e.g., security implications and solutions)?
- Using adaptive links, multi path, and multi-band links. Multi-connectivity.
- Etc., etc. (to be added to a refined)

5. IEEE 802 Standards Supporting Low Latency Communications

The following IEEE 802 standards and amendments can assist or realize in achieving low latency (some in tandem with high reliability) communication.

5.1 IEEE 802 Published Standards with Low Latency features

IEEE 802.1 TSN Family of Standards

IEEE Std 802.1Q-2020: Bridges and Bridged Networks

IEEE Std 802.1AB-2016: Station and Media Access Control Connectivity Discovery (specifies the Link Layer Discovery Protocol (LLDP))

IEEE Std 802.1AS-2020: Timing and Synchronization for Time-Sensitive Applications

IEEE Std 802.1AX-2020: Link Aggregation

IEEE Std 802.1BA-2021: Audio Video Bridging (AVB) Systems

IEEE Std 802.1CB-2017: Frame Replication and Elimination for Reliability

IEEE Std 802.1CM-2018: Time-Sensitive Networking for Fronthaul (summary page)

IEEE Std 802.1CS-2020: Link-local Registration Protocol (approved draft standard)

802.3br Interspersing Express Traffic provides a fundamental latency reduction capability by allowing a large frame to be suspended, transmit a small latency sensitive frame, then resume the suspended frame.

802.11ai Fast Initial Link Setup, 802.11r Fast Handover (“Fast” is a relative term)

IEEE 802.11ax-2021 Enhancements for High Efficiency WLAN

The IEEE 802.11ax amendment was approved February 21, 2021. The amendment improves the performance of Wi-Fi networks in dense areas.

IEEE 802.11ax is designed to operate in 2.4 GHz, 5 GHz, and the newly opened 6 GHz bands. Through increased link efficiency in frequency domain, time domain, and modulation schemes, IEEE 802.11ax can achieve as high as 12.01 Gbps under ideal conditions [6].

Latency is reduced through the use of OFDMA for uplink and downlink, with the associated scheduling by the AP. The use of Multi-User Multi-Input/Multi-Output (MU-MIMO) is extended to the uplink, and the use of 1024 quadrature amplitude modulation (1024-QAM) is enabled to carry more bits per symbol.

1.1.1 802.11ad and 802.11ay (60 GHz)

802.11ad was the first 60 GHz standard, and it defined a scheduled MAC layer. The follow-on IEEE 802.11ay was approved in 2021 and achieves a maximum throughput of at least 20 Gbps using the unlicensed mm-Wave (60 GHz) band, while maintaining or improving the power efficiency per STA.

IEEE 802.11ay can provide a high throughput utilizing various technologies, such as channel bonding/aggregation, MIMO (multiple-input and multiple output), and multiple channel access, etc. [6].

1.1.2 802.11be Extremely High Throughput

IEEE 802.11be is primarily focused on increased data rates, but some of the enhancements also improve latency. Multi-Link Operation (MLO) allows STAs to operate on multiple channels with a single logical connection. MLO can support a single-radio or multi-radio implementation and can reduce latency by transmitting on the first available channel. The introduction of Restricted Target Wake Time (R-TWT) also improves latency by requiring other STA's transmissions to end before the start of the TWT Service Period advertised by the AP.

1.1.3 802.11bd V2X

Low latency is a requirement for V2V use cases. IEEE 802.11bd improves on 802.11p by increasing throughput and implementing PHY adaptations to better support high speed movement (doppler and rapidly changing channel conditions). Latency reduction is primarily achieved by the higher rate, and lower packet loss (and thus retries) from the PHY improvements.

802.15.3 supports low latency, isochronous streaming. Two-way streaming. 802.15.3 specifies fast link setup and teardown. (and future with THz developments)

802.15.4 TSCH (provides more predictable, but not extremely low latency – 100 mS range)

802.15.4z UWB and 802.15.4ab for AR/VR to provide low-latency positioning and low latency audio.

802.16 and 802.22 provide scheduled MAC with predictable latency (10s of mS) Operation in licensed spectrum provides more predictable packet deliver and thus latency, compared to unlicensed, due to the lower potential for interference.

6. Adaptions and Recommendations for IEEE 802 Standards to Enhance Low Latency Communications Support

The 802.1 TSN TG will continue to provide the overall framework and architecture for low latency across multiple standards.

The RTA TIG in 802.11 discussed multiple real-time applications in several domains (gaming, industrial automation, drone control, etc.) and their requirements are summarized in Table 6-1. Real-time applications have been evolving, so do their communication requirements. While voice and video accounted for most of the real-time traffic in the past, new and emerging applications such as real-time gaming, AR/VR, robotics and industrial automation are expected to become more prevalent in the future. Some of these applications also impose new worst-case latency and reliability requirements for Wi-Fi systems. Therefore, one of the recommendations of the RTA TIG to the 802.11 working group is to consider a broader range of real-time application requirements as summarized in Table 6.1. [4]

Use cases		Intra BSS latency/ ms	Jitter variance /ms [4]	Packet loss	Data rate/ Mbps
Real-time gaming [4]		< 5	< 2	< 0.1 %	< 1
Cloud gaming [4]		< 10	< 2	Near-lossless	< 0.1 (Reverse link) > 5Mbps (Forward link)
Real-time video [4]		< 3 ~ 10	< 1~ 2.5	Near-lossless	100 ~ 28,000
Robotics and industrial automation [2][1]	Equipment control	< 1 ~ 10	< 0.2~2	Near-lossless	< 1
	Human safety	< 1~ 10	< 0.2 ~ 2	Near-lossless	< 1
	Haptic technology	<1~5	<0.2~2	Lossless	<1
	Drone control	<100	<10	Lossless	<1 >100 with video

Table 6-1 Requirements metrics of RTA use cases

New capabilities to support real time applications

[1] There may be other wireless applications in industrial automation that are not considered real-time, therefore they are out of the scope of this report.

Potential enhancements and new capabilities to address requirements of emerging real-time applications can be grouped in the following categories:

Extensions of TSN capabilities to 802.11: As described earlier, 802.1 TSN standards are addressing real-time applications over Ethernet and extensions of TSN over 802.11 can help better support such applications over wireless medium. TSN features have already been enabled in 802.11, including traffic/stream identification, time synchronization, and integration with Ethernet bridging. But new extensions are required to address the worst-case latency problems in current Wi-Fi deployments. Time-Aware shaping and redundancy through dual links (FRE capability) are examples discussed in this report, which exist in Ethernet TSN, but need support from 802.11 in other to be adapted to wireless medium as discussed in [7]. Other TSN features may also be considered, such as alignment with the TSN management model defined by the 802.1Qcc standard.

Multiband operation simultaneously: Due to the diversity demands for Wi-Fi networks, dual-band even tri-band AP and STA products have been brought up to market and more features are expected, since nowadays one end user tend to utilize multiple media thus multiple traffic streams. So, requests for high concurrency, reducing impact of interference and traffic differentiation are becoming universal demands. Multiband operation is defined in 802.11be.

Multiband operations simultaneously can benefit not only real-time applications but also those applications request high throughput and traffic separation. [4]

New MAC/PHY capabilities that reduce latency and improve reliability: There is also need for improvements in the 802.11 MAC and PHY layers to enable more predictable latency, which is a fundamental requirement for most real-time application, as discussed previously in the report. It should be noted that for many real-time applications, predicable worst cast latency does not necessarily mean extremely low latency, but the ability to provide more predictable performance is the main requirement. However, in some use cases, the worst-case latency requirement may also need to be low. Another related are for improved identified is reliability. Enabling features that can be used to improve overall reliability of 802.11 links are also needed to support emerging real-time applications. Although operation is unlicensed spectrum makes it difficult to provide hard performance guarantees, many Wi-Fi deployments can be managed. Therefore, it is important to enable capabilities that can be leveraged in managed environments to provide more predictable performance.

Potential areas for further enhancements include: reduced PHY overhead, predictable and efficient medium access, better support for time-sensitive small packet transmissions, improving management and time-sensitive data coexistence, coordination between APs, more flexible OFDMA resource allocation scheme, etc. [4]

These enhancements will be considered in the 802.11 Ultra High Reliability (UHR) Study Group, which will become the 802.11bn Task Group.

7. Conclusion

IEEE 802 standards are addressing low latency requirements on a number of fronts.

Many vertical applications require low latency, both in absolute time, as well as predictability and bounded delivery time.

Wired and wireless media are inherently different. The dedicated nature of the wired medium allows for better control of latency.

The wireless standards operating in unlicensed spectrum have progressed significantly from their early versions in terms of minimizing and managing latency. Progress continues in this area.

Wireless standards are optimized for specific use case and applications. Most of the IEEE 802 wireless standards are trying to reduce latency. To a more limited extent, they are adopting aspects of IEEE 802.1 TSN to further improve latency predictability. The predominate use of unlicensed spectrum by IEEE 802 wireless standards adds to the challenge of delivering predictable, low latency services.

The different IEEE 802 wireless standards address this challenge in different ways: predictive channel access, multiple spatial streams, coordinated multi point transmission, and other new innovations continue to be discussed. Low latency represents a rich area for new innovations and technical approaches.

References

[1] IEEE 802.24 and 802.1 White Paper "Utility Applications of Time Sensitive Networking" https://ieeexplore.ieee.org/document/8870295

[2] IEEE 802.11 TGay Use Cases: https://mentor.ieee.org/802.11/dcn/15/11- 15-0625-07-00ay-ieee-802-11-tgay-usage-scenarios.pptx.

[3] IMT-Vision – Framework and overall objectives of the future development of IMT for 2020 and beyond, Recommendation ITU-R M.2083-0, Sep. 2015

[4] RTA TIG Report : https://mentor.ieee.org/802.11/dcn/18/11-18-2009-06-0rta-rta-report-draft.docx

[5] Goal of the ‘Network Enablers for seamless HMD based VR Content Service’ SG : https://mentor.ieee.org/802.21/dcn/18/21-18-0065-00-0000-21-18-0065-00-0000-goal-of-the-network-enablers-for-seamless-hmd-based-vr-content-service-sg.pptx
[6] Andrew Radford, ‘Why Is Low Latency Important? 5 Key Areas’, *Securus Communications*, 2021, https://securuscomms.co.uk/why-is-low-latency-important/ (December 16th, 2021)
[7] Ali B. Jaber; Mehdi E. Manaa, ‘A Robust Fog-Computing Security Approach for IoT Healthcare Applications’, *IEEE Xplore*, (26 October 2021)