Radio communication for Communications-Based Train Control (CBTC): A tutorial and survey

Farooq, Jahanzeb; Soler, José

Published in:
IEEE Communications Surveys & Tutorials

Link to article, DOI:
10.1109/COMST.2017.2661384

Publication date:
2017

Document Version
Peer reviewed version

Citation (APA):

General rights
Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

- Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
- You may not further distribute the material or use it for any profit-making activity or commercial gain
- You may freely distribute the URL identifying the publication in the public portal

If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.
Radio communication for Communications-Based Train Control (CBTC): A tutorial and survey

Jahanez Farooq, Member, IEEE, and José Soler, Member, IEEE

Abstract—Over the last decade, railway industry has seen a huge transition from conventional railway signalling systems to modern, communication-based signalling systems. Communications-Based Train Control (CBTC) is a modern communication-based system that uses radio communication to transfer timely and accurate train control information. CBTC is the choice of mass-transit railway operators today, with over a hundred systems currently installed worldwide. The safety-related, time-critical applications such as train control impose stringent reliability and availability requirements on the radio communication technology used. IEEE 802.11 Wi-Fi, despite being originally developed for stationary users within a limited area, has prevailed as the de-facto radio technology for CBTC. Unfortunately, very limited literature is publicly available on this topic due to the highly competitive nature of the railway industry. We believe that this paper fills the much-needed gap. It aims to present a comprehensive tutorial, as well as a survey of the state-of-the-art, of CBTC and the role of radio communication in it. The operation and fundamental components of a CBTC system are discussed. A summary of the evolution of the communication technologies used for modern railway signalling is presented. The benefits and drawbacks of using a radio communication technology, particularly Wi-Fi, and the challenges it introduces, are discussed. Best practices in the design of a CBTC radio network and the measures to optimize its availability are discussed, while using the currently in-progress Copenhagen S-line CBTC project as a reference. An overview of the CBTC standardization efforts, as well as the IEEE CBTC standard—frequently overlooked due to its limited scope—is included. The paper is concluded by providing a number of potential directions for future work.

Index Terms—Railway signalling, rail transport, communication-based train control, CBTC, radio communication, Radio Communication System (RCS), Wi-Fi, Wireless LAN, IEEE 802.11.

I. INTRODUCTION

Over the last decade, there has been a huge focus on rail transport due to the reasons including environmental awareness, increased urbanization, population growth, and it being a more energy-efficient, safer, higher capacity, and higher speed transport alternative. Recent studies [1]–[3] show that the European rail market grew from 122 billion euro per year to approximately 150 billion euro in the period 2008–2013, and is expected to grow to approximately 176 billion euro by 2017. Furthermore, it is estimated that a total of 1,077.8 km of rail tracks for the modern, communication-based signalling system CBTC will be installed in the period 2011–2021, compared to only 188.9 km in 2001–2010 [4].

Rail traffic is characterized by poor braking capabilities because of low friction on rails, fixed path, and the inability to avoid obstacles. Therefore, at its most basic, the objective of a railway signalling system (or train control system) is to prevent trains from colliding and derailing [5].

Conventional railway signalling is based on color light signals and train detection with the help of track circuits and axle counters. However this technology is nearly half a century old. It is nearing its expiry in most of the installations worldwide and is responsible for most of the delays experienced every day. This is one reason why the conventional signalling systems are rapidly being replaced by modern signalling systems [2], [4], [6]–[8].

In modern, communication-based railway signalling, different means of telecommunication are used to transfer train control information between the train and the wayside. However, today the term is used almost exclusively for radio-communication-based signalling. CBTC is a modern, radio-communication-based signalling system. Using radio communication, it enables high resolution and real-time train control information, which increases the line capacity by safely reducing the distance (headway) between trains travelling on the same line, and minimizes the numbers of trackside equipment [9]–[12]. CBTC is the first choice of railway operators for mass-transit operations today, with currently over one hundred CBTC systems installed worldwide [9]. Note that although communication-based train control is a generic term, today the term CBTC is used specifically to imply systems used for mass-transit, mostly employing IEEE 802.11 Wireless LAN (WLAN) [13] for radio communication. Thus, CBTC systems are considered distinct from the European Rail Traffic Management System (ERTMS)—another modern, communication-based signalling system, targeted towards mainline railway operations.

Unlike many other research and development areas, the state-of-the-art in CBTC is driven by the industry rather than the academia. In addition, due to the highly competitive nature of the industry, the amount of publicly available literature on this topic that openly discusses implementation details is highly insufficient. The main contribution of this paper is to provide a comprehensive tutorial as well as a survey of the state-of-the-art of radio communication in CBTC. The available industrial and scientific literature on this topic was consulted for this purpose, besides the knowledge acquired from the author’s own experience of working on the devel-

J. Farooq is with Siemens A/S, 2750 Ballerup, Denmark. He is also with DTU Fotonik, Department of Photonics Engineering, Technical University of Denmark, 2800 Kgs. Lyngby, Denmark (e-mail: jahanez.farooq@siemens.com).

J. Soler is with DTU Fotonik, Department of Photonics Engineering, Technical University of Denmark, 2800 Kgs. Lyngby, Denmark (e-mail: joss@fotonik.dtu.dk).
opment of a CBTC system. Denmark is currently one of the front runners in Europe as it is carrying out a total renewal of its entire railway signalling before 2021, with an investment of 3.2 billion euro. This includes the Copenhagen mass-transit network S-train, which will be equipped with a CBTC system [14]. The new signalling system is expected to enable greater capacity and 80% reduction in signalling related train delays [7]. The paper aims for a pragmatic approach, occasionally using the Copenhagen S-train project as a reference. Nonetheless, the information provided is generic in nature and is not restricted to any specific project or supplier.

CBTC systems have historically used IEEE 802.11 WLAN, popularly known as Wi-Fi, as the radio technology, mainly due to its cost-effectiveness. In contrast to radio communication for non-safety related rail applications such as CCTV and onboard Internet, radio communication for safety-related application such as train control imposes stringent reliability and availability requirements. This paper discusses the historical reasons behind the success of Wi-Fi as the de-facto technology for CBTC, despite its lack of support for mobility and susceptibility to interference. It presents the best practices in the design and architecture of a CBTC radio communication network, and the measures to ensuring high system performance.

There has been a general lack of standardization efforts for CBTC, the result of which is that nearly all existing CBTC installations are incompatible, proprietary systems [10]. Although there exists an IEEE standard for CBTC [15], [16], it has not gained much attention from CBTC suppliers due to its limited scope. This paper intends to bring more attention to it.

The rest of this paper is laid out as follows. Section II presents an overview of CBTC, its function and components, and the role of radio communication in it. Section III discusses the evolution of communication technologies for communication-based railway signalling. Section IV presents an overview of the benefits and drawbacks of using a radio technology, particularly Wi-Fi, for CBTC. Roaming, being an inevitable reality in the CBTC systems, is discussed in Section V. Section VI discusses the design and architecture of a CBTC radio network, both onboard and trackside network. Section VII discusses some relevant IEEE 802.11 parameters to optimize the radio communication performance. Section VIII presents an overview of the CBTC standardization efforts, along with a summary of the IEEE CBTC standard. Section IX presents a brief overview of the leading CBTC solutions and suppliers. Section X presents a summary of the future research directions. Finally, section XI concludes the paper.

II. CBTC OPERATION

In CBTC, continuous, high capacity radio communication is used to exchange train control information between the train and the wayside, enabling automatic train control (ATC) functions, namely automatic train protection (ATP) and automatic train operation (ATO).

The train continuously sends its current speed, direction, and location to the wayside over the radio connection. Based on this information received from all trains currently on the track, as well as a train’s braking capability, the traffic control center at the wayside calculates the maximum speed and distance the train is permitted to travel, collectively known as “limit of movement authority” (LMA), and sends it to the train. Based on this information, the train onboard ATC equipment continuously adjusts the train speed and maintains the safety distance to any preceding trains. Thanks to this real-time information exchange, the trackside equipment used in conventional systems, such as color light signals and track circuits, is not needed, and can be removed.

The speed and location of a train is determined using a combination of devices such as speedometers, tachometers, transponders (“balises”), Doppler radar, odometers, and geolocation systems such as Global Positioning System (GPS) [17]. Location accuracy, in particular, is highly critical. Transponders or balises are fixed reference points mounted between rails. As a train passes over a balise, the location information is transmitted from the balise to the train using an antenna mounted under the train. Between the balises, location is continuously estimated using onboard odometry measurements. Any inaccuracies accumulated over distance are corrected when train passes the next balise [17]. The IEEE CBTC standard discussed in subsequent sections recommends a location accuracy of 5 to 10 meters [15]. There are a number of problems associated with using a geolocation system such as GPS as the primary means for localization. The location accuracy of geolocation systems might not be high enough, e.g. to differentiate trains traveling closely to each other. Satellite signals cannot be reliably received inside tunnels. Furthermore, CBTC suppliers are generally reluctant to depend on a system that is controlled by an external authority. Therefore, the use of a geolocation system in CBTC is normally supplementary.

A. Fixed block vs. moving block

In conventional railway signalling, tracks are divided into blocks (or “track sections”), and track circuits are installed to determine if a train is inside a block. Each block is protected by a signal. Various factors dictate the length of the block, including how busy the line is, the maximum allowed speed on that line, the maximum speed and braking capabilities of different trains, sighting, etc. When a train is inside a block, since there is no real-time method to determine its exact location inside the block, the entire block is declared as occupied, and other trains are not permitted to enter it. As the boundaries of these blocks are fixed, regardless of a particular train’s speed and braking capability, and are further reinforced by track circuits, this type of operation is called “fixed block operation” [17], [18].

In contrast, in the “moving block operation” employed in CBTC, thanks to the real-time communication between the train and the wayside, the train location is continuously updated. As a result, the occupancy zone—or block—“moves” with the train and reflects its actual location. There are no fixed blocks boundaries. As shown in Fig. 1, this allows trains to run closer to each other.
control approach to quantitatively describe the effects of communication failures on train control performance resulting in an information gap—the difference between the received and the actual state of the front train. The work in [26] optimizes IEEE 802.11 MAC layer parameters—see Section VII—to minimize the energy consumption caused by unplanned braking as a result of unacceptably large information gap.

C. Data traffic requirements

The typical size of a CBTC control message is 400-500 bytes. A message transmission time of shorter than 100 milliseconds is normally supported. Given that the typical frequency of these messages is about 100-600 milliseconds, data requirement for a CBTC system is typically in the range of 20-40 kbps, and not more than 100 kbps [19], [27]–[33].

D. Components and networks

This section discusses the major components of a typical CBTC system, as well as the two-way communication network that connects the train and the wayside. This network further consists of the following three integrated networks: [20], [34], [35]

1) Train onboard network
2) Train-to-trackside radio network
3) Trackside backbone network

The train onboard network and the trackside backbone network use Ethernet, while the train-to-trackside radio network generally uses Wi-Fi.

1) Onboard components: This section discusses the major onboard components of a CBTC system, as shown in Fig. 2. Together, these components comprise the train onboard network.

a) Vehicle On-Board Controller/Computer (VOBC): The onboard equipment includes Vehicle On-Board Controller/Computer (VOBC), sometimes also called Carborne Controller or Onboard Control Unit (OBCU). This system is responsible for sending train control information to the wayside on periodic basis. It either includes, or works together with, the onboard ATP and ATO subsystems [15].

b) Onboard ATP and ATO: The ATP and ATO subsystems are part of the onboard ATC functionality. ATP controls safety-related functions and ATO controls the actual train driving functions. Each of these has both onboard and wayside components.
As probably the most critical subsystem, the ATP subsystem helps prevent collisions as a result of the driver’s failure to observe a signal or speed restriction. It monitors and controls the train speed and applies brakes if necessary. The ATO subsystem is responsible for automating the train operation, including basic operations normally performed by a driver, such as starting and stopping the train, energy-efficient braking and acceleration, and stopping accuracy.

c) Radio Communication System (RCS): Another critical onboard component is RCS, or Data Communication System (DCS). RCS is typically a combination of software and hardware, including radios and antennas, and is responsible for the radio communication between the train and the wayside. RCS can either be a completely independent system or integrated into VOBC. If independent, the computer system running RCS is also frequently referred to as a Train Unit (TU).

2) Wayside components: Fig. 3 illustrates typical wayside components of a CBTC system. The terms wayside and trackside are often used interchangeably. However, trackside generally contains the components located either on or close to the tracks, and is considered a part of the wayside.

A Zone Controller (ZC), or Wayside Controller, is responsible for controlling a particular zone in the railway network. Dividing the wayside network into multiple, independent zones, such that each zone comprises its own wayside infrastructure, improves availability even if one or more zones experience failures. The fundamental function of a ZC is to maintain safe train separation in its zone. A ZC also typically includes the wayside ATP and ATO subsystems [9, 15].

Fig. 3. CBTC wayside components

The ATP subsystem of a ZC manages all the communication with the trains in its zone. It is also this subsystem that calculates the movement authority for every train in its zone. A Computer-based Interlocking (CI) system is either included as an independent system or as a part of the ATP subsystem. CI controls the trackside equipment such as point/switch machines and signals, and is responsible for setting routes for trains. The ATO subsystem provides all the trains in its zone with their destination as well as dwell times [9].

Independent from the ZC is the automatic train supervision (ATS) system, which is responsible for monitoring and scheduling the traffic.

Trackside is divided into multiple Wi-Fi cells, each served by one Access Point (AP). Fig. 3 uses the green and red colors to differentiate the APs’ radio coverage areas. In the later sections of this paper, they will be used to represent two different radio frequencies as well. APs are either deployed on one side of the track or both, in alternating fashion. Trains communicate to the APs through a radio connection. This constitutes a typical CBTC train-to-trackside radio network. APs are in turn connected to the wayside components through the trackside backbone network.

A typical configuration of the trackside backbone network is star-topology, as shown in Fig. 4 (a), where each AP is connected directly to the wayside infrastructure using fiber optic cables [34], [36]–[41].

Fig. 4. Star vs. Ring based trackside network

An advanced alternative is ring-topology, shown in Fig. 4 (b) [42], [43]. This configuration minimizes cabling, as the distance between an AP and the backbone network is usually much larger than the distance between two adjacent APs. An inherent limitation of a ring-based network is that a single failed node can disrupt the whole network. However, a number of Ethernet ring redundancy protocols, such as Media Redundancy Protocol (MRP), exist to mitigate this problem [44], [45]. Additionally, multiple rings can be employed to enable excessive redundancy, or to keep the number of nodes in a ring network under the limit.

III. THE EVOLUTION OF COMMUNICATION TECHNOLOGIES FOR RAILWAY SIGNALLING

A. Conventional signalling systems

In a slightly evolved form of the conventional signalling, known as "track-based train control" (TBTC), rails are used for communication between train and wayside infrastructure. These systems use coded alternating current (AC) track circuits, also called audio frequency (AF) track circuits because of the range of the frequencies used, to modulate data [17], [46]. The train control data sent via rails is then used for cab-signalling—the feature of displaying signal aspect information to the driver inside the train—, and to enforce the permitted speed [18]. Since in these systems, track circuits are used to determine train location, this sort of signalling is also referred to as "track-circuit signalling". This technology can be considered an early form of communication-based train control. An immediate example is the driverless Copenhagen Metro.

However, the low resolution of train location determined by track circuits and the low capacity of rail communication leads to less accurate train location information. This results in larger headways to ensure safety and thus low line capacity.
Typical headway in conventional train control systems is several minutes [5], [20].

B. Modern signalling systems

1) Inductive loop-based CBTC: Early CBTC systems in the 1980s used inductive loop as the communication technology [5], [47]. This type of signalling was also called "transmission-based train control" (TBC) [9]. The first CBTC system was based on this technology, and was installed in Toronto, Canada, in 1985, on the Toronto Transit Commission Scarborough RT Line [5], [48]–[50]. Two other such early inductive loop-based systems were called VAL, deployed on Lille light metro in 1983, and Meteor, in service on Paris Line 14 since 1998 [51]–[53].

In these systems, inductive loop cables were mounted on tracks, and were coded with certain frequencies at regular intervals. The train verified its location by reading these signals via a detector mounted beneath it [17], [54]–[56]. As seen, this method could be considered an advanced alternative to the track-circuit signalling discussed above. For this reason, track-circuit signalling is occasionally argued to be a form of communication-based train control.

In contrast to today’s CBTC systems that work in the GHz frequency range with Wi-Fi, inductive loop systems worked in the kHz range. However, despite its demerits, inductive loop is a proven technology that has been used for railways for three decades. It is cost-effective, as it uses unshielded standard wire, which is easy to repair. However, the downside is that it is not easy to install, and is vulnerable to vandalism and theft [49].

2) Radio-based CBTC: As discussed above, the modern CBTC systems use continuous and high capacity radio communication between the train and the wayside infrastructure to transmit train control information. The high resolution and highly accurate train location enables the "moving block" operation. The result is short headways and increased line capacity. A typical headway in CBTC systems is 90 seconds or less [5], [15], [20], [21]. Furthermore, it enables advanced features such as driverless and unattended train operations [18], [57].

The first radio-based CBTC system was supplied by Bombardier and was installed at San Francisco airport in 2003 [9].

Radio-based CBTC systems can roughly be divided into two categories: those based on the modern, high capacity radio communication—which can be further divided into custom and commercial-off-the-shelf (COTS) technologies—and those based on the older leaky waveguide technology.

a) Custom and Commercial-off-the-shelf (COTS) radio: Most of the earlier CBTC radio systems were custom solutions, built specifically to fit a project’s requirements, and used a proprietary radio technology. An example is Andrew Corporation’s Model 2400 radio solution, based on the leaky waveguide technology, used by Bombardier in its initial CBTC installations. Its cost was $22,000 (per radio), roughly 100 times the cost of a Wi-Fi based solution [49]. The downside of custom solutions is their lack of compatibility with systems developed by other suppliers. Later, Bombardier opted for a spread-spectrum based COTS solution developed by Safetran Systems Corporation (now Siemens) [53], which cost only $1600, though still roughly 10 times the cost of a Wi-Fi based solution [49].

Similarly, to keep the radio system independent of a particular supplier, New York City Transit (NYCT) opted for a COTS solution for its Culver and Canarsie Line projects. The chosen solution was called RailPath, and was developed by Springboard Wireless Networks Inc. It was based on the spread-spectrum technology and operated in 2.4 GHz band [53], [58], [59].

The radio system used in the Copenhagen S-train CBTC system, called Airlink [42], is based on Wi-Fi. However, previous generations of Airlink still use proprietary custom-built radio technology, based on spread-spectrum and operating in the 2.4 and 5.9 GHz bands. The latest project using this spread-spectrum based system is the recently contracted NYCT’s Queens Boulevard Line [60].

b) Leaky waveguide: A leaky waveguide is a coaxial cable with periodic openings in its shielding to allow radio signals leak out or in, thus acting as a continuous antenna. Leaky waveguide is also known as leaky feeder, leaky cable, or radiating cable. For decades, it has been successfully used to provide voice radio service in metros [49], [61].

Leaky waveguide offers certain advantages. Radio communication in open-air locations is unpredictable in general, as the propagation loss a signal experiences depends heavily on the obstructions it encounters in its way. Leaky waveguide involves very limited open-air communication, which takes place over a very short distance—normally in the range of 0.3 to 0.6 meters—between the leaky cable and the receiver antenna on the train. Thus, leaky waveguide guarantees a more predictable propagation loss and is less susceptible to interference [47], [62]–[64].

Given these advantages, certain railway operators have used a combination of radio communication and leaky waveguide. While radio communication is used in tunnels, leaky waveguide is used in open-air locations where interference is significantly higher, or in the critical locations where radio communication is exceedingly problematic [62]–[64]. An example is stations in tunnels, where several standing trains could obstruct the line-of-sight (LOS) path to the nearest AP. However, one challenge in these solutions is the seamless switching between the two technologies at the transit areas. A separate set of antennas must be used for each technology [63].

The downside with leaky waveguides is that they are not cost-effective and installation and maintenance requires lots of effort, especially in the congested tunnel environments [19], [63]. Furthermore, when installed in open-air locations, they are prone to signal degradation due to environmental effects such as rain and snow. For these reasons, leaky waveguide has not been proven very popular for CBTC systems [61].

IV. Why Radio/Wi-Fi for CBTC?

This section presents the major reasons radio, particularly Wi-Fi, has been chosen as the communication technology for
CBTC, starting by summarizing its benefits and drawbacks first.

A. Benefits

The major benefits of radio-based CBTC systems include [5], [50]:
- High capacity — i.e. data throughput
- Low costs and easy upgradability — cost-effective radio equipment, easy software upgrade
- Less trackside equipment — e.g. as a result of removal of track circuits and axle counters
- Easy scalability — e.g. by adding more radio equipment
- Easy installation and maintenance — as a result of fewer cables
- Fault-tolerance/redundancy — e.g. through multiple radios, and overlapping radio coverage
- Low susceptibility to vandalism — as a result of fewer cables

In addition, the major reasons for choosing Wi-Fi as the radio technology include [39], [49], [50], [65]:
- Freely available Industrial, Scientific and Medical (ISM) frequency band
- A large vendor market and industry support
- Low costs — cost-effective and readily available COTS radio equipment
- Interoperability among multiple vendors, thanks to the Wi-Fi Alliance
- Open standard protocols

Contrary to mainline railway (i.e. long-distance, suburban trains), the number of trains in mass-transit is larger, also as a result of shorter headways. Therefore, mass-transit requires a higher capacity radio technology compared to GSM-R (Global System for Mobile Communications - Railway), adopted by the European mainline standard ERTMS [47].

While actual CBTC traffic itself does not demand high data rates, as discussed above in Section II-C, it is still the key to enabling modern CBTC applications such as remote diagnostics and maintenance, remote software upgrade, CCTV, transmission redundancy, as well as passenger infotainment applications such as onboard Internet. Data rate for supporting these applications can easily reach a few megabits per second per train. Wi-Fi, supporting data rates of up to 300 Mbps, not only fulfills this requirement but also has the capability to do so in the near future.

The underground nature of the mass-transit railway means it is not feasible to install antennas on high masts like in open-air locations. It is therefore inevitable to install numerous APs along the track to cover a large area. The availability of low cost COTS Wi-Fi equipment therefore has played a decisive role in the success of Wi-Fi for CBTC [10].

It is worth pointing out that although significantly less compared to a conventional technology such as inductive loop, Wi-Fi equipment is still vulnerable to vandalism, nonetheless. In particular, the trackside APs, while enclosed in protective metal enclosures and mounted on masts, are still visible and in reach. Common examples are causing damage to the AP enclosure, and cutting the cables to the enclosure or to the AP antennas.

B. Drawbacks

A few of the drawbacks of choosing Wi-Fi include:
- Susceptibility to interference
- Requires stringent security measures
- Lack of support for mobility
- Short range
- Network congestion

Susceptibility to interference from other Wi-Fi and non-Wi-Fi users is a known issue (see next section). However, again, it proves to be less of a problem due to the underground nature of the mass-transit transport. In underground environments, the probability of interference from other users is comparatively lower and can be controlled more effectively [10].

Although the security concerns outlined here apply to any broadcast-based radio technology, the use of ISM band (see next section) makes them even more relevant to Wi-Fi. Appropriate security measures are required to be in place to prevent unauthorized users from connecting to a CBTC AP with their Wi-Fi devices, sniffing the traffic, or stealing bandwidth resources. Of relevant concern are the jamming attacks that can disrupt the entire radio network, or the man-in-the-middle attacks where an intruder may pose as a legitimate CBTC AP, causing trains to connect to it. Authentication and end-to-end data encryption methods thus are highly critical. Therefore, relevant standards specifying appropriate security measures, such as EN 50159 discussed in Section VIII, are normally implemented by CBTC systems.

An intrusion detection system in the CBTC context is proposed in [66] that, by using an advanced authentication protocol, prevents intrusion attacks by increasing the computational complexity required to perform such attacks.

The IEEE 802.11 standard was primarily developed to replace cables in local area networks such as office environments, and therefore inherently does not support mobility and large ranges [20]. Handover was therefore not considered. For this reason, the CBTC radio communication systems generally implement their own handover algorithms [28]. The generally low speeds of mass-transit trains further minimize this inherent lack of support for mobility in IEEE 802.11 [10].

In cellular networks such as GSM (Global System for Mobile communication) or LTE (Long-Term Evolution), the distance between a mobile node and a base station is normally large. Comparatively, in mass-transit, the distance between a train and a trackside AP is short, mostly due to the congested tunnel environments. This makes the short range of Wi-Fi less an issue [10]. The problem is further minimized by having large number of APs deployed on the trackside.

Poor Quality of Service (QoS) due to congestion in contention-based medium access networks such as IEEE 802.11 is a well-known issue, especially when the number of users is large. However, it is not as serious an issue in the CBTC scenario. It is unlikely that there is more than one train in a Wi-Fi cell at a time, because trains on rails cannot get too close to each other for safety reasons. This is due to the larger
length of train compared to the size of a cell. Furthermore, in a typical configuration, only two radios are transmitting, one at each end of the train. The probability of both ends being in the same cell are therefore further decreased [28], [40], [41].

C. Frequency band and interference

Nearly all CBTC installation today work in one of the three, license-free ISM bands: 900 MHz, 2.4 GHz, and 5 GHz. Of these, 2.4 GHz is the most popular among CBTC suppliers, followed by 5 GHz [61]. Table I lists the ISM band frequency ranges together with their user applications [67]–[72].

<table>
<thead>
<tr>
<th>Frequency range</th>
<th>Users</th>
</tr>
</thead>
<tbody>
<tr>
<td>902 - 928 MHz</td>
<td>Microwave ovens, cordless phones, industrial heaters, military radar, RFID, IEEE 802.11ah</td>
</tr>
<tr>
<td>2.4 - 2.4835 GHz</td>
<td>IEEE 802.11b/g, microwave ovens, cordless phones, Bluetooth, garage doors openers, baby monitors, car alarms, printers, keyboards/mice</td>
</tr>
<tr>
<td>5.725 - 5.825 GHz</td>
<td>IEEE 802.11a/h</td>
</tr>
<tr>
<td>61 - 61.5 GHz</td>
<td>IEEE 802.11ad</td>
</tr>
</tbody>
</table>

1) Interference: Interference, both co-channel and adjacent-channel, is a well-known issue in Wi-Fi networks.

As discussed above, one major reason for choosing Wi-Fi is its use of the ISM band. This means railway operators don’t have to worry about acquiring a license from a regulatory body.

In the US, these band has been designated by the Federal Communications Commission (FCC) as license-free, which means it can be used by anyone, without the need for acquiring a license. Some restrictions on the transmission power do apply though [73]. In Europe, similar regulations are applied by European Telecommunications Standards Institute (ETSI), Body of European Regulators for Electronic Communications (BEREC), and European Commission (EC).

However, since the number of users using a license-free band is significantly larger, there is a higher probability of interference from other users in the band. As an example, in a recent study [74] on Chongqing Rail Transit Line 1 in China, up to 1,300 unique SSIDs were observed over a period of one second—SSID (Service Set Identifier) is a sequence of characters that uniquely identify a Wi-Fi network (or AP). The increasing use of the 2.4 GHz band for CBTC systems by railway operators has therefore raised concerns. As an example, CBTC failures at Shenzhen Metro have been attributed to interference caused by the non-CBTC Wi-Fi users in the surrounding locations [75], [76]. These incidents are studied in [77] together with a few interference mitigation methods. Given these reasons, an RF (radio frequency) site survey is normally first conducted to determine the amount of interference before planning AP placement. However, the rapid and widespread proliferation of smartphones and other handheld devices means it is not trivial for such surveys to accurately predict the interference even in the near future.

To minimize adjacent-channel interference, adjacent APs in CBTC systems are deployed on alternating frequency channels. This is discussed in greater details in subsequent sections.

2) Licensing: Acquiring a licensed band is the optimal solution to prevent the risk of interference in CBTC systems. However, it is a lengthy administrative process with limited chances of success due to the scarcity of spectrum. The spectrum that is available exists in bands for which there is little or no radio equipment available. Allocation in these bands therefore would require a significant investment in research and development by radio vendors prior to deploying a fully functional CBTC system [61], [73].

There are a few exceptions though, notably the Copenhagen S-train CBTC system, for which the 5.925-5.975 GHz band has been licensed.

3) Factors for choosing a frequency band: In CBTC, choosing between 2.4 GHz and 5 GHz frequency bands is generally driven by the following factors:

- Availability of cost-effective radio equipment: Such equipment is more likely to be available at 2.4 GHz because of the large vendor market as discussed above. CBTC vendors will rather provide a communication system based on readily available COTS equipment than developing their own proprietary solution [61].

- User density and interference: As discussed above, due to a significantly large number of Wi-Fi and non-Wi-Fi users, the 2.4 GHz band is much more prone to interference compared to the 5 GHz band, as seen in Table I.

- Signal range: In general, the higher the frequency of a radio wave, the shorter the distance it can travel. Thus, the 2.4 GHz radio waves cover a substantially larger distance than the 5 GHz waves, with the same transmission power. This is due to the characteristics of high frequency radio waves that not only attenuate faster but also do not penetrate solid objects nearly as well as the low frequency waves. However, an advantage of high frequency signals is that since they do not travel as far, they also interfere less with the neighboring signals.

- Ease of installation: Operating frequency also drives the number of APs installed. The shorter signal range of 5 GHz radio waves means shorter distances between APs, resulting in a greater number of AP installations. Furthermore, frequency also drives the location and height of AP installation, due to the propagation characteristics discussed in Section VI-A.

- Number of available channels: When configuring frequency channels for adjacent APs, as a rule of thumb, non-overlapping channels are preferred to further limit the interference. However, only 3 and 4 non-overlapping channels are available in IEEE 802.11b (DSSS modulation) and IEEE 802.11g (OFDM modulation), respectively [78], [79]. Comparatively, IEEE 802.11a, which operates in the 5 GHz band, enables 23 non-overlapping channels [69].

V. ROAMING IN CBTC SYSTEMS

In contrast to cellular communication, roaming in railway environments is not a mere possibility but is an inevitable reality. Even worse, unlike the cellular networks, Wi-Fi are
short range networks, where larger networks are built by deploying more APs closely together. This means APs are placed at regular intervals on the trackside network, such that their coverage areas overlap, and a train has to continuously find a new suitable AP and re-connect as it moves along.

A critical aspect of roaming in CBTC thus is how a radio communication system smoothly switches from one AP to another (i.e. handover), without causing interruptions and delays in the communication. As discussed above, a large handover latency might result in a delayed reception of the movement authority information, and the train might have to apply emergency brakes [15], [20].

Furthermore, since in CBTC it is very common to deploy adjacent APs on different frequencies, the radio communication system must switch between them when switching from one AP to another. This, combined with the high speeds of modern trains, results in rapidly changing channels, and renders the handover algorithms successfully used in the stationary Wi-Fi environments, inefficient for CBTC [28], [79].

A. Handover frequency and latency

The frequency of handover is determined by the distance between the two APs (or the AP coverage areas, see Section VI-A) and train speed. The handover in IEEE 802.11 is the so-called "hard handover", in which the mobile node breaks the current connection before establishing the next connection, resulting in delays and packet loss [20], [80]. High speed and short inter-AP distance result in more frequent handovers, further worsening the situation [41], [81]. As an example, a train travelling at 180 km/h with APs deployed at every 300 meters will experience handover every 6 seconds.

Studies show that the number of packets lost due to handover is much larger than that due to radio propagation [28]. The authors in [20] propose a method for determining packet loss rate based on the handover time, the AP coverage range, and the overlapping coverage area between APs. They show that with a train speed of 200 km/h, the maximum handover time of 180 milliseconds, and the overlap area of 20 meters, the calculated packet loss rate is approximated to be 10%.

Handover time in CBTC is typically in the range of 70-120 milliseconds, with 1 second as an upper limit [20]. As long as this time is shorter than the CBTC control message interval discussed above, it does not impose a serious threat, as it only means one lost message in the worst case.

B. Roaming algorithm

Normally a smooth transition is achieved by equipping a train with at least two radios, one at each end, such that at least one of these radios is always connected to an AP.

In its simplest form, it works as follows. As the train moves, the front radio continues to search for a new AP. When it finds a new AP, it breaks the current connection and establishes a new one with the new AP (connection 1 in Fig. 5), while the rear radio stays connected. Next, the rear radio switches the connection to the new AP (connection 2), while the front radio stays connected. Sophisticated roaming algorithms might develop some sort of a distributed algorithm to prevent both radios from roaming at the same time. Next, the first step is repeated and the front radio connects to a new AP (connection 3). An analogy is that of climbing a rope—or more accurately—a Tarzan-style swing from rope to rope.

A similar multi-radio based roaming method for CBTC proposed in [82] aims for a transport layer solution rather than the data link layer to further minimize the handover latency—to very close to zero. It uses Stream Control Transmission Protocol (SCTP), and benefits from its multi-homing feature to establish simultaneous connections to two APs.

However, before the execution of a handover, it must be detected, i.e. when to execute the handover. How this is achieved is not specified by the IEEE 802.11 standard, and therefore, CBTC systems typically develop their own roaming algorithm. Typical approaches are to monitor the quality of the link e.g. by monitoring the number of un-acknowledged packets (i.e. packet loss), or by monitoring when signal quality falls below a certain threshold [83]. A couple of approaches for the latter are to measure signal quality (e.g. by means of Received Signal Strength Indicator (RSSI)) from the beacon frames received from an AP, and to measure how consist the signal quality is, by monitoring the frame reception rate [29]. If more than one potential APs are found for the new connection, the same criteria is used to select the best AP among them. The highly dynamic environments of rail transport makes the detection of handover further challenging.

1) Roaming threshold: A typical approach in CBTC systems is to perform the handover as soon as the train receives a signal from a new AP, with the power above a certain threshold, even if the signal power of the current AP is greater and still increasing—note that the threshold is only applicable if there is more than one AP to choose from. This is illustrated in Fig. 6, where, for example, a handover from AP-1 to AP-2 is performed at a time when the current signal power of AP-1 (red) is greater than that of AP-2 (green).

The objective mainly is to avoid the acute drop in the signal power as the train moves past the current AP. This is in part due to the misconception that directional antennas don’t have coverage at their backside, as implied in Fig. 6 as well. However, this is far from reality. The "front-to-back ratio" antenna parameter specifies the ratio of radiations transmitted in the forward direction to that transmitted in the

![Fig. 5. Roaming/handover in CBTC](http://dx.doi.org/10.1109/COMST.2017.2661384)
backward direction [84]. For example, the HUBER+SUHNER
Sencity SPOT-S antenna used in Copenhagen S-train CBTC
has a front-to-back ratio of 20 dB [35]. This means, its
coverage in backward direction is -20 dB worse than that in
forward direction, but there is still coverage, except in the
rare cases when it is entirely blocked by the mast on which
the antenna is installed. This misconception leads to incorrect
implementations of roaming algorithms, which connect to a
new AP prematurely, anticipating an acute signal drop as the
train moves past the AP.

Yet another approach is to use two different thresholds: a
"leaving threshold" and a "joining threshold". The roaming is
performed if the current AP's signal power falls below the
leaving threshold and/or the new AP’s signal power is above
the joining threshold.

Nonetheless, these thresholds must be set carefully. A too
low leaving threshold may result in a prolonged connection to
the current AP. The result is a delayed roaming, which may
lead to the train losing the signal altogether as it moves past
the AP. A too low joining threshold can result in the train
connecting to an AP with poor signal quality [50].

Note that if APs employ directional antennas—see Sec-
tion VI for a detailed overview of different antenna
configurations—the roaming performance also depends on the
roaming direction. In Fig. 7 (a), the direction of the train
movement is the same as the antenna pointing direction on
the APs. Thus, it allows sufficient time for the train to see
the gradual decrease in signal strength and connect to a
new AP based on the leaving threshold. In contrast, when
roaming in the opposite—face-to-face—direction, as shown in
Fig. 7 (b), the train sees a gradual increase in signal strength.
Nevertheless, it still needs to connect to the next AP based
on the joining threshold, before it moves past the current AP.
Otherwise, it might hold the connection to the current AP for
too long and then disconnect abruptly [27], [50].

C. IEEE 802.11 handover

This section presents a brief discussion on the IEEE 802.11
handover mechanism, with the intention of highlighting po-
tential improvements.

Handover in IEEE 802.11 has 3 phases: (1) scanning, (2)
authentication, and (3) re-association.

1) Scanning: Scanning is the process of finding a suitable
AP to connect to. Of the three handover phases, scanning takes
the longest time. As per studies, the latency of the scanning
phase accounts for approximately 90% of the total handover
latency [20], [85].

In passive scanning, a wireless node waits for the beacon
messages sent by the APs to announce their presence. Passive
scanning is slow as most hardware vendors set the beacon
interval to 100 milliseconds, by default. In active scanning, the
node sends probe request messages, to which nearby APs reply
with probe response messages. Generally, in CBTC systems,
active scanning is adopted to minimize the latency [20], [28].
However, since normally the node needs to probe all frequency
channels (11 in 802.11b, for example), it still takes significant
time [28].

The choice of frequency channels to use when deploying
a trackside radio network is often independent of the
development of the CBTC radio communication system. As
a consequence, even if only two channels are used, it is not
uncommon that a CBTC radio communication system is
developed in a way to still scan all channels by default, to
be on the safe side. One reason is that often these CBTC
systems are either not customized for a particular customer’s
needs, or, are unaware of the actual channels used. In an
adaptive approach, once having learned the channels, only
these channels are used afterwards, only to fall back to all
channels in case of failure to find an AP.

Different approaches are taken to reduce the time spent
on scanning. Often a sort of background passive scanning is
employed where the node learns about the next available APs
while still connected to the current AP [86]. The authors in
[36] propose a similar solution, called "frequency combination
algorithm", which actively scans for as many as four neigh-
boring APs as the train moves to create a "neighbor graph".
Such an algorithm is capable of providing redundancy in the
event of more than one AP failure.

Due to the linear nature of a trackside radio network, the
next AP to connect to can be pre-determined, provided that
the train maintains an up-to-date database of AP information,
and, all APs are in a healthy state.

In [87], authors propose a location-based solution, in which
a node knowing its own position and the direction of move-
ment, receives information about neighbor APs from a location
server. The server also performs proactive authentication and
association for the node, further reducing the handover latency.
Another solution, presented in [88] uses a GPS server to collect
the location of the node and determines which AP is best suited
for handover. A related location-based solution specifically
developed for CBTC is presented in [28].

2) Authentication: In the authentication phase, a node es-
blishes its identity with the AP it found in the scanning
phase, by exchanging special authentication messages. However, depending on the vendor implementation, this phase could take significant time, e.g. up to one second if an IEEE 802.1X [89] based centralized security architecture is used that involves communicating with an authentication server, such as RADIUS [90].

A few Wi-Fi products support the IEEE 802.11f [91] extension, also known as Inter-Access Point Protocol (IAPP). IAPP offers a pre-authentication method in which an AP, upon successful authentication with a node, shares the node’s authentication information with the nearby APs [28]. The authentication process can thus be skipped when connecting to those APs subsequently.

In CBTC scenarios, where seamless handover is critical, one approach is to skip the authentication phase altogether. However, the drawback is that authentication then must be performed at packet level using a higher layer security protocol, such as IP Security (IPSec).

3) Association: In the association phase, the node registers itself to the AP by exchanging special messages, so that the AP could forward data to/from it. There is no room for improvement in this phase’s latency as it is dictated only by the message transfer delay [28].

VI. RADIO NETWORK CONFIGURATION

This section presents the best practices and some of the key parameters considered while designing a CBTC train-to-trackside radio network.

A. Inter-AP distance

Redundancy is critical to providing reliable radio communication in CBTC and is a deciding factor when planning the number and placement of the trackside APs, as well as the onboard radios. Redundancy is further discussed in Section VI-C1. As the train’s movement is fixed, the configuration of the train-to-trackside radio network is linear, which is helpful in reducing installation efforts. The APs are placed as close to the track as possible to get the best possible signal quality on the train and to avoid any obstructions in the line-of-sight (LOS) path. To provide continuous connectivity, the inter-AP distance, which is the distance between two adjacent APs, is chosen in a way that APs’ coverage areas overlap. RF (radio frequency) link budget calculations are typically made to determine the inter-AP distance, and AP signal range plays a key role in these calculations [31]. As stated above, an RF site survey is normally subsequently performed to determine the number and placement of APs.

A common approach is to use a short inter-AP distance, as well as a high transmission power, to overcome interference from other devices/users.

To aid a smooth handover, the inter-AP distance is designed in a way that the train is always inside the coverage of at least two APs. Another advantage of a short inter-AP distance is that the front radio can hear not just the AP ahead of it, but also the next one. It increases the availability as the front radio has twice as many APs as it needs [61]. Note that the adjacent APs must be deployed on different frequencies to avoid interference in this case.

The following subsections discuss parameters affecting the inter-AP distance.

1) AP signal range: Table II lists the "rule-of-thumb" ranges for the indoor and outdoor environments for IEEE 802.11b, which offers the largest ranges compared to the other most commonly used IEEE 802.11 standards, 802.11a and 802.11g [92].

<table>
<thead>
<tr>
<th>Mode</th>
<th>Modulation</th>
<th>Outdoor range (m)</th>
<th>Indoor range (m)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 Mbps</td>
<td>DSSS</td>
<td>550</td>
<td>50</td>
</tr>
<tr>
<td>2 Mbps</td>
<td>DSSS</td>
<td>388</td>
<td>40</td>
</tr>
<tr>
<td>5.5 Mbps</td>
<td>CCK</td>
<td>235</td>
<td>30</td>
</tr>
<tr>
<td>11 Mbps</td>
<td>CCK</td>
<td>166</td>
<td>24</td>
</tr>
<tr>
<td>5.5 Mbps</td>
<td>PBCC</td>
<td>351</td>
<td>38</td>
</tr>
<tr>
<td>11 Mbps</td>
<td>PBCC</td>
<td>248</td>
<td>31</td>
</tr>
<tr>
<td>6 Mbps</td>
<td>OFDM</td>
<td>300</td>
<td>35</td>
</tr>
<tr>
<td>12 Mbps</td>
<td>OFDM</td>
<td>211</td>
<td>28</td>
</tr>
<tr>
<td>18 Mbps</td>
<td>OFDM</td>
<td>155</td>
<td>23</td>
</tr>
<tr>
<td>24 Mbps</td>
<td>OFDM</td>
<td>103</td>
<td>18</td>
</tr>
<tr>
<td>36 Mbps</td>
<td>OFDM</td>
<td>72</td>
<td>15</td>
</tr>
<tr>
<td>48 Mbps</td>
<td>OFDM</td>
<td>45</td>
<td>11</td>
</tr>
<tr>
<td>54 Mbps</td>
<td>OFDM</td>
<td>36</td>
<td>10</td>
</tr>
</tbody>
</table>

However, the range of IEEE 802.11 radio signal depends on various factors and can be enhanced.

The parameters such as the antenna height, transmission power, gain, and receiver sensitivity can be adjusted to enhance the signal range. Additionally, the signal range depends on the operating frequency, discussed in Section IV-C3, and the propagation loss. When planning a CBTC radio network, all these parameters are used in the link budget calculations to determine the AP coverage. Note that, the aim with these calculations is often to provide a guaranteed, minimum number. To be on the safe side, often the worst-case propagation loss, as well as various "margins", e.g. "link margin" and "fade margin", are used in these calculations. As a result, a mere +6 dBm difference in the actual received power doubles the achievable distance. As an example, a study [93] found that the APs deployed for the Copenhagen S-train CBTC could be heard as long as 4 kilometers away, despite the link budget calculation of approximately 600 meters.

Note that in CBTC environments, due to the limitations of track and tunnel structures, antennas are generally installed at lower heights, potentially limiting their range [27].

Additionally, significantly longer distances of up to 100 kilometers in point-to-point links can be achieved by adjusting the IEEE 802.11 MAC (Medium Access Control) layer parameters such as ACK (acknowledgement packet) timeout, slot time, and Contention Window (CW) size [94], [95]. See Section VII for further details.

The above stated methods have widely been used to enable deployment of low-cost, long-distance Wi-Fi based wireless networks in rural areas. These kind of networks are formally known as "Wi-Fi over Long Distance" (WiLD) [94].
2) **Receiver sensitivity**: Receiver sensitivity is the minimum signal power required at the receiver antenna to demodulate the signal. The more advance the modulation scheme used is, the greater the signal power (or Signal-to-Noise Ratio (SNR)) required to correctly demodulate the signal [84].

Table III lists the receiver sensitivity requirements specified by the IEEE 802.11a standard, which is based on OFDM and operates at 5 GHz [13]. When making link budget calculations, it is recommended that the receiver sensitivity values specified by the hardware manufacturer are followed.

<table>
<thead>
<tr>
<th>Data rate</th>
<th>Modulation</th>
<th>Coding rate</th>
<th>Minimum sensitivity (dBm)</th>
</tr>
</thead>
<tbody>
<tr>
<td>6 Mbps</td>
<td>BPSK</td>
<td>1/2</td>
<td>-82 dBm</td>
</tr>
<tr>
<td>9 Mbps</td>
<td>BPSK</td>
<td>3/4</td>
<td>-81 dBm</td>
</tr>
<tr>
<td>12 Mbps</td>
<td>QPSK</td>
<td>1/2</td>
<td>-79 dBm</td>
</tr>
<tr>
<td>18 Mbps</td>
<td>QPSK</td>
<td>3/4</td>
<td>-77 dBm</td>
</tr>
<tr>
<td>24 Mbps</td>
<td>16-QAM</td>
<td>1/2</td>
<td>-74 dBm</td>
</tr>
<tr>
<td>36 Mbps</td>
<td>16-QAM</td>
<td>3/4</td>
<td>-70 dBm</td>
</tr>
<tr>
<td>48 Mbps</td>
<td>64-QAM</td>
<td>2/3</td>
<td>-66 dBm</td>
</tr>
<tr>
<td>54 Mbps</td>
<td>64-QAM</td>
<td>3/4</td>
<td>-65 dBm</td>
</tr>
</tbody>
</table>

Receiver sensitivity is directly related to a particular Bit Error Rate (BER), Packet Error Rate (PER), or Frame Error Rate (FER) [84]. As an example, the IEEE 802.11a standard states that the minimum required receiver performance at the 54 Mbps data rate is -65 dBm with a PER of 10% or less, as seen in Table III.

As discussed above, AP signals can travel significantly large distances in favorable propagation conditions. However, since radio signals attenuate greatly due to various propagation phenomena, to be on the safe side, an effective range of only 200-300 meters is assumed in CBTC. The studies in [20] and [27] show that in poor propagation conditions, the probability that the received power falls below the receiver sensitivity increases when an inter-AP distance greater than 200 meters is used. Furthermore, the probability of receiving a signal of acceptable power level is greater than 95% when an inter-AP distance of smaller than 200 meters is used. Choosing a distance shorter than 200 meters, on the other hand, means higher costs as well as more frequent handovers.

For these reasons, typical inter-AP distances range from 100 to 600 meters, depending on the track and terrain topology, e.g. curves, elevations, slopes, obstructions, etc., and the transmission power used [39]. A study of CBTC installations show that an inter-AP distance of 200-300 meters is more common, however greater distances of more than 350 meters have also been seen [20], [27]-[29], [34], [37], [63], [96]. One example is the Copenhagen S-train CBTC where an inter-AP distance of approximately 600 meters has been used.

3) **Radio propagation modeling**: As a signal travels from a transmitter to a receiver, it incurs loss in signal power due to various propagation phenomena such as reflection, refraction, diffraction, absorption, and multipath effect, due to the environment and the obstructions in the way. Propagation loss is further dependent on the height and location of the antennas, and the distance between the transmitter and the receiver [97].

High train speed results in rapid Doppler shift, leading to serious degradation of the signal [97], [98]. Furthermore, fading characteristics of high-speed trains significantly differ from those of the cellular communication environments where the classical Hata and COST-231 Walfish-Ikegami models have successfully been used to predict propagation characteristics. In contrast, the WINNER II model [99] defines propagation calculations more relevant to high-speed suburban and urban scenarios [98].

In [100], authors adapt the well-known Okumura-Hata model to railway operations. The adapted model takes into account the presence of water in close proximity of railway tracks, such as banks and shores, resulting in additional reflections and diffractions over water. However, the model is more relevant to long-distance railway operations, where terrains with water presence are more common, than it is to CBTC.

Additionally, in contrast to the cellular networks, the low AP antennas heights in CBTC result in Fresnel zone limiting the propagation of high frequency radio waves to short ranges [27], [82]. Studies such as [101] derive models to more closely represent propagation loss in these scenarios.

An alternative method could be to extract propagation model from field data. For example, the RSSI information retrieved from packets received at the train and the trackside APs could be used to determine propagation characteristics of the environment. A challenge, however, is to identify packets lost due to interference rather than poor signal quality. Similar methods have been applied to the vehicle-to-vehicle (V2V) environments with considerable success [102], [103]. However, its application in the CBTC context has yet to be seen.

The works in [104] and [105] propose propagation models for CBTC systems based on leaky waveguide.

a) **Radio propagation in tunnels**: Radio communication in tunnels is highly critical as most mass-transit trains run underground [106]. Even though interference from other users is relatively limited in tunnels, they still present a challenging environment for radio propagation. Propagation loss in tunnels depends on signal frequency, tunnel dimensions and shape, curves in the tunnel, construction material and surface roughness of the tunnel walls, as well as the structures the tunnel contains. Reflection is a major contributor of the signal attenuation, besides the other aforementioned propagation phenomena [32], [61], [106], [107]. Normally, the more complex the structures contained in the tunnel, and the greater their roughness, the greater the propagation loss [107]. Tunnels with turns/curves cause additional reflections, where the propagation loss increases with the decreasing radius of the curve [106], [107]. The signal attenuation due to the above phenomena is further emphasized at higher frequencies. All this highlights the reasons why leaky waveguide, discussed above in Section III-B2b, is still occasionally preferred over radio communication for tunnels.

Furthermore, curves block the line-of-sight (LOS) signal between transmitter and receiver, a problem that is normally
solved by increasing AP density around curves.  

Modeling propagation loss in tunnels is therefore more complex compared to open-air locations, and advanced propagation models are continuously being developed [29], [106]–[108]. Part of the problem is that carrying propagation loss measurements in tunnels is problematic, as they cause inconvenience for normal traffic [107]. Various methods are employed to accurately model propagation loss in tunnels, including numerical methods for solving Maxwell equations, waveguide models, ray-tracing models, and two-slop models [106].

Nevertheless, there are a number of limitations with the use of the above stated methods. Most of these methods are based on mathematical computations. Detailed modeling of every object in a tunnel, whose length can be several kilometers, might involve large computational complexity and significant amount of time [107]. Furthermore, these methods do not always accurately predict loss in tunnels with highly dynamic environments, i.e. with moving trains and people. Moving trains cause further signal attenuation as a result of reflections from the train, and the "shadowing effect" when train passes the AP [32], [107]. Additionally, the predicted loss is less accurate if the environment where the model is applied differs from that against which the model was developed.

Tunnels with limited curves and objects demonstrate favorable propagation environments, and act as giant waveguides. A waveguide is a device, usually a hollow metal tube, which transports radio waves from one point to another with minimum propagation loss, a phenomenon known as "the waveguide effect" [61], [62], [97], [106], [107].

In CBTC, APs are mostly installed either on tunnel walls or very close to them. However, studies show that this presents the worst propagation conditions, mainly due to the reflections from the walls, and that the optimal position of the APs is in the middle of the tunnel instead [107].

### B. Antenna configuration

An omni-directional antenna provides equal coverage in all directions, resulting in a wider coverage area, though the covered distance is short. In contrast, a directional (or uni-directional) antenna provides coverage in a specific direction, resulting in a larger distance but a narrow coverage area.

The type of antenna used varies across CBTC solutions.

1) **Trackside antenna:** In general, directional antennas provide better coverage in the line-of-sight (LOS) environments and omni-directional antennas perform better in the non-line-of-sight (NLOS) environments, for example in tunnels with curves [107].

The use of omni-directional antennas on linear environments such as a trackside network provides a more "wide" coverage. Due to their short range, the train should only "see" one AP ahead of it, if the APs are appropriately spaced. This means faster and less complex roaming as the "AP selection" part of the algorithm can be avoided [50]. The obvious disadvantages are a large number of APs required to cover a given area, and higher susceptibility to interference from nearby users, e.g. Wi-Fi hotspots at train stations.

Often two directional antennas are used instead, facing in opposite directions, as discussed subsequently in Section VI-C1. It reduces the number of APs and makes for a more predictable, linear pattern, which suits well to a trackside network. It does offer some challenges, however. The width of the coverage area of a directional antenna depends on the size of the antenna’s beamwidth. An AP antenna with a very narrow beamwidth means the train cannot see the AP before it is properly "aligned" to the AP’s coverage area [50].

Both with directional and omni-directional antennas, the train sees a gradual increase in the signal strength as it approaches the AP, followed by a slight dip while it is adjacent to the AP. However, where it differs is when the train subsequently moves away from the AP. In the case of an omni-directional antenna, the train sees a gradual decrease in the signal strength. In contrast, in the case of a directional antenna, it sees a rapid drop in the signal strength immediately after moving past the AP [50].

Depending on the track and terrain topology, APs are mounted on masts to provide optimal coverage as well as ease of maintenance [39]. The height of these masts ranges from 50 centimeters to the more typical 4 meters for underground installations, and 4-6 meters for open-air installation [19], [20], [27], [63], [109]. Often the antenna height is chosen to be reasonably above the level of the train roof, e.g. between 0.5 and 1 meter, to ensure that radio waves are not shielded by the train. Ray-tracing models can be employed to determine the desired height.

2) **Onboard antenna:** Antennas are generally installed at a sufficient height on the train roof so that the line-of-sight (LOS) path to the AP does not get obstructed by other trains. Generally directional antennas are preferred onboard, for the reason discussed above. However, using directional antennas on both AP and train might make roaming harder, as it means that their coverage areas have to be aligned perfectly to be able to see each other. For this reason, some CBTC systems use a combination of directional and omni-directional antennas, e.g. directional antennas on trains and omni-directional antennas on APs, or vice-versa [30], [96].

### C. System availability

Since it is about railway operations and the safety of passengers, system availability is highly critical. It is no surprise that CBTC suppliers boast of 99.999% (6.05 seconds of downtime per week) or better system availability of their solutions, particularly of the radio communication systems [64], [110], [111].

1) **Redundancy:** In CBTC, redundancy is the key to high availability. The general rule is that at any given location on the track, minimum two APs shall be available to connect to. A train is typically equipped with two TUs (Train Unit), one at each end, to provide sufficient redundancy. Each TU is typically equipped with one radio, though solutions with two radios are also seen. Each radio is then equipped with one or two antennas.

As discussed in detail subsequently, redundancy is realized by the following various means:

- Two TUs per train
- Two radios per TU
Two antennas per radio
Two or more frequencies
Redundant AP coverage areas
Redundant APs per location
Redundant AP/trackside backbone networks

Redundancy is additionally ensured by diversity. Having two antennas per radio, separated by a certain distance, provides spatial diversity, as two independent signals can be received at a given time, of which the stronger signal can be used. Antenna diversity helps overcome multipath fading [50], [97]. Additional spatial diversity is automatically provided by having two TUs per train, separated by the length of the train [47]. Employing two or more frequencies, as presented later in this section, is itself a form of diversity—called frequency diversity—as it decreases the probability that all signals in a particular area are corrupted in the same way. Likewise, a typical practice in CBTC of repeating the transmission of the same information, e.g. over different radios/antennas, is a form of temporal diversity [50], [97]. Yet another form of redundancy is the power redundancy. When using the multiplexing topology discussed in Section II-D2, to make the AP rings (or backbone networks) completely independent of each other, they are deployed with independent power supplies.

The following discussion presents an overview of some of the typical configurations employed in CBTC. In Fig. 8 and Fig. 9, presented for this purpose, the green and red colors of the AP coverage areas represent two different frequencies. Furthermore, the onboard antennas do not represent a specific antenna type, i.e. directional or omni-directional, unless specifically stated.

Fig. 8 (a) shows a configuration with no onboard redundancy. In this configuration, there is only one TU on the train, with one (directional or omni-directional) antenna. The trackside AP also has one radio with one directional antenna. Radio frequencies have been used in an alternating fashion. For example, an inter-AP distance of 300 meters implies that the distance between the two consecutive APs operating on the same frequency is 600 meters, which helps in minimizing the interference.

Note that on APs, the use of directional antennas only in one direction as in Fig. 8 (a) might lead to the famous "hidden node problem", where two nodes are in the range of a common node.
but not in each other’s range. This makes the "carrier sensing" protocol CSMA/CA (Carrier Sense Multiple Access/Collision Avoidance) used to avoid collisions in IEEE 802.11 MAC ineffective, as the two nodes cannot hear each other. As discussed above in Section VI-A1, it is not uncommon that the actual AP signal range is much larger than the "guaranteed" range. Let’s suppose that in Fig. 8 (a), AP 3’s signal can be heard by a train currently in AP 1’s coverage area. AP 1 and AP 3 cannot hear each other, because (1) AP 1’s antenna is pointed in the opposite direction, and (2) it has a very low front-to-back ratio. This may lead to a situation where AP 1 starts transmitting, while AP 3 is already transmitting, resulting in a collision. This problem is solved in configuration Fig. 8 (c) where AP 1 has an additional antenna in the opposite direction, allowing it to hear AP 3’s transmission, and thus suspend its transmission.

Fig. 8 (b-e) presents configurations with onboard redundancy. Onboard redundancy is provided by having two independent TUs on the train, such that these TUs are connected to two different onboard networks, as seen earlier in Fig. 2. The specific configuration varies across solutions. In most configurations, both TUs are connected to APs all the time, where the second TU is either used to transmit simultaneously, as a fallback, or to connect to the next AP in advance. In certain solutions, TUs are purposely configured with different SSIDs so that they do not connect to the same AP.

A typical configuration, presented in Fig. 8 (b), is to equip each AP with one radio and one omni-directional antenna. In contrast, the configuration shown in Fig. 8 (c) uses two directional antennas pointing in each direction, which extends the radio coverage. Compared to Fig. 8 (b), there are half as many APs in this configuration to provide the same coverage area. The antenna configuration in Fig. 8 (c) closely resembles that of the Copenhagen S-train CBTC system, which, with a radio coverage area of approximately 600 meters at each side, allows a distance of approximately 1200 meters between the two consecutive APs operating on the same frequency.

Fig. 8 (d) shows a configuration in which each AP is equipped with two radios, enabling it to operate on two frequencies/SSIDs simultaneously. If TUs are configured with different SSIDs, this configuration is more suitable compared to the one in Fig. 8 (c), in which the coverage area of a single AP might be large enough to cover the whole train, thus making it less likely for the two TUs to connect to different APs.

The hidden node problem discussed above appears here again. Since an AP’s radios/antennas are on two different frequencies, AP 1 is still not able to hear what AP 2 transmits on the "green" frequency. Fig. 8 (e) shows an alternative configuration that solves this problem by employing frequencies in the "ABBA" fashion, rather than the normal "ABAB" fashion. The placement of the same frequency antennas face-to-face improves the effectiveness of the CSMA/CA protocol. Additionally, the increased distance between the new potential hidden nodes, e.g. AP 1 and AP 3 (not visible in the figure), minimizes the probability of AP 3’s signals reaching AP 1.

Fig. 9 presents configurations with additional wayside redundancy. As shown in Fig. 9 (a), besides the basic wayside redundancy in the form of overlapping coverage areas, additional redundancy is provided by deploying two separate AP backbone networks, and placing APs in the two networks in alternating fashion [41], [80]. The failure of one network thus does not affect the other. Note that this very much resembles the multi-ring topology discussed in Section II-D2.

The configuration in Fig. 9 (b) adds two additional levels of redundancy: (1) improved coverage redundancy, as there is a greater overlap between the coverage areas of the neighboring APs, and, (2) AP redundancy, as there are two APs at each location. Even a complete failure of one of the backbone networks won’t affect the coverage. Nonetheless, this additional level of redundancy comes at the expense of an increased—twice as many—number of APs.

If the two TUs are configured to connect to APs in different backbone networks, as shown in Fig. 9 (b), the result is less frequent handovers, as handover is required only when both TUs lose connections. Note that when two APs are placed next to each other in this fashion, their antennas can point to the same direction, as in Fig. 9 (b), or in opposite directions, outward or inward, as in Fig. 9 (c) and Fig. 9 (d) [41], [80].

As discussed before, normally directional antennas are believed to be highly directional, with no coverage at their backsides. The objective behind installing antennas pointing inward seen in some configurations [80], as in Fig. 9 (c), is to avoid this blind spot. However, as discussed in Section V-B, directional antennas are not that directional in real life. Thus, instead, the coverage behind antenna looks more like as shown in Fig. 9 (d).

Fig. 9 (e) shows a configuration with a complete coverage redundancy. A close examination of the figure shows that there is a near 100% overlap of the coverage areas of the neighboring APs—half of the coverage areas of AP 1-1 and AP 1-2 (green) are completely hidden by the coverage area of AP 2-1 (red). Note that this is fundamentally the same configuration presented earlier in Fig. 8 (c), except that the inter-AP distance has been reduced greatly to enable this near complete overlap. The Copenhagen S-train CBTC system uses a similar coverage area overlap.

Finally, Fig. 9 (f) presents the same configuration as in Fig. 9 (b), except that it enables additional onboard redundancy by using two radios/antennas per TU instead of one, allowing four simultaneous connections at a time.

A study carried out in [80] compares the availability of the configuration with no onboard redundancy, presented in Fig. 8 (a), to that of two configurations with onboard redundancy. The first configuration with onboard redundancy is similar to the configuration in Fig. 9 (b). There are two APs at each location and the train is in the coverage of one such pair of APs at any time.

The second configuration with redundancy is similar to the configuration in Fig. 8 (b). There is only one AP at each location, but the inter-AP distance is halved to make sure that each TU is in the coverage of two APs. A TU connects to the AP offering the better signal quality, and switches to the other AP in the event of failure. As a train is always in the coverage of four APs, the communication will be interrupted only if all four connections fail.
The study shows that the unavailability of the configuration with no onboard redundancy is more than 2% with a train speed of 120 km/h. Compared to that, both configurations with onboard redundancy decrease the unavailability to below $10^{-4}$. Furthermore, since the second configuration with onboard redundancy makes two additional AP connections available at any given time, it offers better availability than the first, despite experiencing more frequent handovers.

In [112], authors present a mechanism that improves availability by enabling train-to-train communication. Hereby, trains act as relays for other trains, by forwarding train control information to the neighboring trains. Not only does it mitigate communication failures, e.g. if an AP breaks down, it also minimizes the communication delays caused by handovers. Likewise, the work in [33] uses Coordinated Multi-point (CoMP) transmission and reception, a recent technology targeted towards improving performance in modern cellular networks. With CoMP, a train can communicate to a cluster of base stations (BS) simultaneously, rather than to only one AP as in conventional CBTC systems. The coordination of multiple BSs improves handover latency and enables spatial diversity by combining multiple received signals.

2) Diagnostics: A highly important feature most CBTC systems provide is diagnostics. In the event of a failure of a wayside or onboard equipment, identifying the failed equipment can be a cumbersome task. This is especially true for the hundreds of APs deployed on the trackside, or equipment on a train en route. Remote diagnostic, data logging, and timely
failure alerts are therefore inevitable features that are required by the IEEE CBTC standard as well, and are provided by nearly all CBTC systems [113]–[118].

Given the number of radio hardware involved, the probability of hardware and/or radio link failure is particularly high. For this reason, a radio communication system typically provides periodic self-testing of radio hardware. This often includes testing of the radio propagation conditions as well, e.g. for excessive radio interference [61], [115].

VII. RELEVANT IEEE 802.11 PARAMETERS

This section presents some of the IEEE 802.11 MAC parameters that can be used to improve the IEEE 802.11 performance in CBTC systems.

A. Slot time and ACK timeout

In the IEEE 802.11 MAC protocol, the values of the following three parameters depend on the distance between the sender and the receiver.

1. Distributed Interframe Space (DIFS): The minimum time a node has to sense the channel idle before transmitting, to avoid collision with a transmission from another node. DIFS is defined as following:

   \[ DIFS = SIFS + 2 \times \text{slot time} \]

2. ACK timeout: After transmitting a packet, the sender waits for an acknowledgement packet (ACK). If not received within the ACK timeout interval, the packet is considered lost, and sender retransmits the packet. ACK timeout is defined as following [13], [94], [119]:

   \[ \text{ACK timeout} = SIFS + \text{slot time} + \text{aPHY-RX-START-Delay} \]

As seen, the value of the above two parameters is based on slot time, defined as following:

3. Slot time: The value of slot time depends on air propagation time, which the standard defines to be 1 µs—the time it takes for radio waves to propagate 300 meters at the speed of light.

Thus, if the distance between the sender and the receiver is greater than 300 meters, the value of the air propagation time parameter must be adjusted accordingly [94]. This was not supported until the introduction of the "Coverage Class" parameter in the recent revisions of the standard [13], [94].

Short Interframe Space (SIFS) is the shortest of the IEEE 802.11 MAC waiting times. It is used to separate transmissions once a node has acquired the medium, e.g. an ACK packet in response of a data packet. This prevents other nodes, who must wait the longer DIFS time, from acquiring the medium meanwhile.

Of the above parameters, ACK timeout is of particular importance, especially when the value of air propagation time has not been adjusted according to the distance. SIFS is consumed at the receiver before sending the ACK packet and aPHY-RX-START-Delay is the time required for the PHY layer to communicate the packet reception to the MAC layer [13]. Thus, it only leaves a slot time for the roundtrip, restricting the maximum distance between the sender and the receiver to 150 meters [94].

The IEEE 802.11 standard does not define a value for ACK timeout and its value varies between manufacturers. Therefore, in CBTC scenarios, where AP range of more than 300 meters is anticipated, its value must be set appropriately. If the value is set to too small, the sender starts retransmitting before an ACK could possibly have been received, or may as well collide with the ACK on its way. If it is set to too large, the sender waits unnecessarily long before retransmitting. Both situations may result in low throughput as well as large delays. As a rule of thumb, for every 300 meter increase in distance above 300 meters, 2 µs must be added to the ACK timeout [27], [94], [95].

Modern Wi-Fi drivers, such as MadWiFi for Linux, provide a means to automatically calculate the slot time and ACK timeout parameters according to the desired distance [94], [120].

The values for SIFS, slot time and aPHY-RX-START-Delay for different IEEE 802.11 standards are listed in Table IV [13].

B. Retransmissions

Given the highly dynamic environment of rail transport, the probability of transmission errors is high, leading to retransmission of packets. The IEEE 802.11 retry limit parameter (i.e. limit on the number of retransmissions) can play an important role. Setting its value to too large results in large delays.
Conversely, a value too small results in a high packet loss [20].

Packet loss has critical impact on system performance as it essentially means train control information cannot be sent to the train in time. Similarly, packet delay must be smaller than the CBTC message interval to make sure that the train control information received is real-time [27].

Additionally, too large retry limit negatively affects the handover latency, as the train may keep associating to the original AP even if its signal quality has fallen below the roaming threshold [20].

The study in [27] shows that the packet loss rate drops dramatically with an increasing number of retransmissions, to as low as $10^{-8}$ with 6 retransmissions. Furthermore, packet delay peaks up to 90 milliseconds, which, though, is still within the limits of the normal CBTC message interval, as discussed in Section II-C.

On the other hand, since CBTC control messages are sent at regular, short intervals, and typically simultaneously on multiple radios/antennas, retransmissions might not be required. In fact, given the real-time nature of the rail transport, retransmissions may even result in larger end-to-end delay and outdated information. One approach, therefore, is to disable retransmissions, and thus ACKs as well.

### C. Packet size

The contention-based medium access of the IEEE 802.11 MAC protocol means small size packets are recommended. Large packets lead to more collisions, resulting in larger end-to-end delays. It is recommended that the packet size not be larger than 1500 bytes in order to satisfy the latency requirements of CBTC. The IEEE CBTC standard specifies a nominal delay of 500 milliseconds to 2 seconds, as a guideline [15].

1) **Coherence time**: The frame error rate (FER) in IEEE 802.11 increases with the increasing train speed, as a result of the Doppler shift. This leads to a decrease in the coherence time of the channel, which is inversely proportional to the Doppler shift [20], [27], [97]. Coherence time is the duration over which the channel can be assumed constant. Thus, frame duration shall ideally be shorter than the coherence time. The results show that the coherence time decreases rapidly with the increasing train speed, and barely fits the frame duration at the speed of 120 km/h. This results in a channel response that fluctuates widely in the duration of the frame [121].

### D. User Datagram Protocol (UDP) as transport protocol

As with other real-time applications, when designing a radio communication system for CBTC, UDP is preferred over Transmission Control Protocol (TCP) as the transport layer protocol [29]. Since CBTC control messages are sent at short intervals, lost, delayed or erroneous packets are generally not critical. The overhead caused by TCP’s handshake and error checking and correction functions can thus be avoided.

### VIII. STANDARDIZATION

The IEEE 1474.1 standard [15], [16], originally published in 1999, defines performance and functional requirements for CBTC. An additional standard 1474.3 [122], published in 2008, defines recommended practice for CBTC system design and functional allocations. However, unlike European Union’s standard for mainline railway operations, ERTMS, the IEEE CBTC standard serves as mere guidelines, and is not strictly followed by the suppliers. As a result, nearly all existing CBTC installations are incompatible, proprietary systems [10].

As an example, of all the CBTC supplier advertisement material consulted for this study [42], [43], [52], [110], [113]–[116], [118], [123]–[132], only Ansaldo STS’s [116] claims to be compliant to the standard.

Additionally, International Electrotechnical Commission (IEC), and its counterpart in Europe, European Committee for Electrotechnical Standardization (CENELEC), are responsible for the development of standards for the rail industry [133]. These standards address both general, safety related, and software related requirements [5], [134]–[141]. Table V lists the relevant standards, with equivalent standards listed next to each other [142], [143].

In the US, American Railway Engineering and Maintenance-of-Way Association (AREMA) is responsible for the development of a manual for recommended practices in railway. Sections 21-23 of this manual address communication-based signalling [144].

European Union’s research project MODURBAN [145] has similar objectives to develop core system architecture and key interfaces for urban guided rail systems.

### A. IEEE CBTC standard

This section outlines a few key and relevant requirements from the IEEE CBTC standard 1474.1.
1) **Definition:** The IEEE standard defines a CBTC system as a continuous, automatic train control system with the following primary characteristics:

- High-resolution train location determination, independent of track circuits
- Continuous, high capacity, bi-directional train-to-wayside data communications
- Train-borne and wayside equipment capable of implementing ATP functions, as well as optional ATO and ATS functions

2) **Performance and functional requirements:** The standard states that in the event of equipment or data communication failure, trains shall continue to move safely, in degraded mode, e.g. at reduced speeds, with the help of a supplementary wayside system (i.e. for train detection).

Additionally, the standard specifies a number of parameters to achieve high level of performance, along with their typical values. **Table VI** lists some of the most relevant parameters.

Equipment reaction times include the time required to calculate new movement authority limit at the wayside after receiving a location update from the train, and the time to determine a new ATP profile on the train after receiving a new movement authority limit.

The standard states that the CBTC equipment shall have a design life of 30 years. Additionally, a CBTC system shall enable, among others, the following to provide for ease of maintenance.

- Maintenance and diagnostic capabilities, including remote diagnostic capabilities
- Built-in test capabilities
- Timely identification of failed components and functions
- Data logging, enabling recreation of the events leading to an error
- Periodic verification of ATP hardware/software/data

The standard specifies parameters for developing a safe braking model, and provides with an example of a typical model as well. The safe braking model must take into account any location inaccuracies, e.g. due to interruptions in the radio communication.

3) **Radio communication requirements:** The standard states that the quality of the radio communication link between the train and the wayside shall be verified periodically.

<table>
<thead>
<tr>
<th>Category</th>
<th>Parameter</th>
<th>Typical value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Performance limitations</td>
<td>Maximum number of trains that can be handled by a Zone Controller</td>
<td>10 to 40 trains</td>
</tr>
<tr>
<td>Location</td>
<td>Onboard train location measurement Resolution Accuracy</td>
<td>± 0.25 m to ± 6.25 m</td>
</tr>
<tr>
<td>Location</td>
<td>Resolution of wayside calculated movement authority limits Accuracy</td>
<td>± 5 m to ± 10 m</td>
</tr>
<tr>
<td>Speed</td>
<td>Onboard speed measurement Resolution Accuracy</td>
<td>± 0.5 km/h to ± 2 km/h</td>
</tr>
<tr>
<td>Speed</td>
<td>Resolution of wayside calculated speed limits Accuracy</td>
<td>± 3 km/h</td>
</tr>
<tr>
<td>Communication delay</td>
<td>Delay in train control messages, in both directions Wayside</td>
<td>0.5 s to 2 s</td>
</tr>
<tr>
<td>Communication delay</td>
<td>Onboard WAYSIDE</td>
<td>0.07 s to 1 s</td>
</tr>
<tr>
<td>Equipment reaction time</td>
<td>Onboard</td>
<td>0.07 s to 0.75 s</td>
</tr>
</tbody>
</table>

The following functional requirements are specified. The communication link shall be able to:

- Support all required ATP, ATO, and ATS functions
- Provide continuous coverage, including in tunnels, cuts, elevated structures, and slopes
- Support bi-directional data transfer with sufficiently low latency
- Support safe, timely, and secure delivery of train control messages

**IX. CBTC PROJECTS AND SOLUTIONS**

Currently over 150 radio-based CBTC projects exist worldwide, including both operational and ongoing projects. Fig. 10 shows a breakdown of these projects according to the regions and suppliers [9], [42], [110], [146]–[150].

Table VII lists the names of the major CBTC solutions as well as their individual components, including the radio communication systems and the frequency bands used [21], [42], [43], [52], [61], [110], [113]–[116], [118], [123]–[132], [151].

GE Transportation and Hitachi are two comparatively young players in the CBTC market. GE Transportation has recently become a part of Alstom [152]. On the other hand, Hitachi now partly owns Ansaldo STS [153].

**X. FUTURE RESEARCH DIRECTIONS**

Following are some of the challenges/improvements anticipated in the future.

**A. Need for more stringent standardization**

As discussed above, the existing CBTC standards are not strictly followed by the suppliers—one reason why different CBTC solutions are highly incompatible. There is a need to put efforts to define a common standard, along the lines of ERTMS, the European Union’s initiative to create a single train control system standard to enable cross-border interoperability.

**B. A common standard for mass-transit and mainline**

CBTC and ERTMS are the two most well-known communication-based railway signalling systems that exist today. However, the two systems are widely incompatible...
when it comes to their wayside infrastructure as well as radio communication technologies.

Mass-transit is characterized by short distances and line capacity, and thus CBTC makes a suitable choice [10]–[12]. In contrast, mainline is characterized by long distances, high speeds and complex networks, and thus use ERTMS as the signalling standard [10], [57], [154], [155].

As discussed before, the use of Wi-Fi as the radio technology incurs significant installation and maintenance costs as it requires a large number of APs to be installed. This, combined with its limited range, susceptibility to interference, and handover issues at high speeds, makes it unsuitable for mainline. Mainline railway operations thus make use of wide area network technologies, e.g. GSM-R for ERTMS.

To ensure interoperability between the CBTC and ERTMS networks, a new trend is to use hybrid train equipment, e.g. the London Crossrail project, which will have trains equipped with both CBTC and ERTMS equipment [10], [11]. However, given that the on-board and wayside infrastructure for the two systems is widely different, such a convergence incurs costs.

For railway operators, managing two incompatible signalling systems is inefficient in terms of cost, scalability and interoperability [10]–[12]. This creates a need to identify the similarities and differences of the two systems and define a common standard. A common system will enable trains with compatible on-board and wayside infrastructure to travel seamlessly between the two networks, using heterogeneous radio technologies based on their availability, e.g. Wi-Fi, GSM-R, LTE, satellite, etc.

Such a common standard is the focus of the recent initiatives such as Next Generation Train Control (NGTC) [156] and shift2rail [157] by the European Commission and the European Rail Industry (UNIFE).

C. Alternative radio communication technologies

1) LTE: LTE has recently been in focus as an alternative technology for ERTMS as well as CBTC [24], [104], [158]–[162]. The high capacity and large coverage it offers, as well as its potentially long life span, makes it a worthwhile alternative to Wi-Fi for CBTC as well. Alstom, Siemens and Ansaldo STS, in collaboration with Huawei and ZTE, respectively, have already taken initiatives to adapt LTE for CBTC [163]–[166].

LTE’s high capacity ensures it can support additional features such as voice communication, passenger Internet, live CCTV video streaming, and Passenger Information Systems. The vehicular communication (V2X) feature in LTE Release 12 and the upcoming releases 13 and 14 has gained a lot of attention from the railway industry. In addition to the already available vehicle-to-infrastructure (V2I) communication, the introduction of the device-to-device (D2D) feature in the upcoming releases will enable vehicle-to-vehicle (V2V) communication. In a CBTC scenario, it can be used to enable direct communication between trains, without the overhead involved in the train-to-wayside communication. This might play an important role in reducing the end-to-end delay, thereby resulting in even shorter headways [167], [168].
2) 802.11p: The IEEE 802.11p [169] standard, also known as Wireless Access in Vehicular Environments (WAVE), like LTE, is also targeted towards vehicular communication, both V2V and V2I. Typical IEEE 802.11p applications include safety-related traffic updates, route guidance, parking assistance etc. [170].

In USA, IEEE 802.11p has been allocated the 75 MHz Dedicated Short Range Communications (DSRC) spectrum in the 5.9 GHz (5.8-5.925 GHz) Intelligent Transportation Systems (ITS) band. The band is divided into 1 control channel and 6 service channels. Similarly, in Europe, it has been allocated a 20 MHz spectrum [170]. The limited number of control channels reduces the time required to search for an AP, thereby making the process of connection establishment faster [83].

The DSRC band is a free but licensed band, in that it can be used free of cost provided a license has been acquired. It is more restrictive in terms of the usage and the radio technologies, and thus automatically reduces most of the interference problems inherent in the ISM band [171]. Table VIII summarizes the data rates as well as communication ranges supported [170].

<table>
<thead>
<tr>
<th>Features</th>
<th>Europe</th>
<th>USA</th>
</tr>
</thead>
<tbody>
<tr>
<td>Radio spectrum</td>
<td>20 MHz</td>
<td>75 MHz</td>
</tr>
<tr>
<td>Data rate</td>
<td>250 Kbps</td>
<td>3.27 Mbps</td>
</tr>
<tr>
<td>Communication range</td>
<td>15-20 m</td>
<td>1000 m</td>
</tr>
<tr>
<td>Radio frequency</td>
<td>5.8 GHz</td>
<td>5.9 GHz</td>
</tr>
</tbody>
</table>

IEEE 802.11p allows all nodes to be on the same channel and configured with the same SSID by default. The "WAVE mode" feature allows nodes to communicate to each other immediately by using the so-called "wildcard SSID", without having to first associate to an AP. Furthermore, an AP uses a special type of beacon to advertise itself. A node can connect to the AP simply by receiving this beacon advertisement. This reduces the overhead associated to a normal IEEE 802.11 handover [171].

A study in [172] evaluates the performance of IEEE 802.11p, using a linear, highway like scenario. The results show that the control channel traffic can successfully be received even at a distance of 2.5 km, and the delay experienced by the control traffic is under 100 milliseconds as long as the traffic load is under 1000 packets per second.

There has been very limited research about the use of IEEE 802.11p in the context of CBTC. The study in [173] concludes that IEEE 802.11p is more suitable for CBTC than IEEE 802.11a, comparing the end-to-end delay and throughput. The study in [82], discussed above, focuses on handover.

3) IEEE 802.11ah: A promising new technology is the upcoming IEEE standard 802.11ah, also known by its marketing name Wi-Fi HaLow. IEEE 802.11ah is supposed to operate in the license-free Sub-1-GHz, 900 MHz ISM band, as shown in Table I. Note that it is the same frequency range as used by the cellular networks today, particularly GSM. Lower frequency implies longer ranges and lower power consumption. Low frequency radio waves are less affected by the various propagation phenomena, and penetrate well through walls and other obstacles. IEEE 802.11ah supports data rates ranging from 150 kbps to 347 Mbps, and introduces enhancements to enable an AP support a very large number of simultaneous nodes [174], [175]. For CBTC, it means a trackside network with a significantly fewer number of APs, and compact and low power radio equipment. Furthermore, a newly introduced coding scheme makes it more robust against Doppler shift, a well-known problem for moving nodes [174].

4) 60 GHz and IEEE 802.11ad: Radio technology operating in the extremely high, millimeter-wave (mm-Wave) frequency range of 60 GHz presents great potential. The 60 GHz frequency band is significantly less crowded compared to the 2.4 and 5 GHz bands, and enables extremely high data rates [176]. This technology gained further momentum with the release of the IEEE 802.11ad standard that operates in the 60 GHz ISM band, and provides data rates of up to 7 Gbps [177]. However, the range offered at these frequencies is short—a few tens of meters—as the high frequency waves are greatly attenuated by the various propagation phenomena discussed above. Nonetheless, larger ranges can be achieved by using high transmission power and high gain antennas [176], [177]. Manufacturers, such as IgniteNet [178], are already introducing equipment capable of providing ranges of up to 1.5 kilometers. Furthermore, due to the short wavelength of the radio waves at these frequencies, they are attenuated by the oxygen molecules in the atmosphere, a phenomenon called "oxygen absorption". To compensate for it, these systems use highly directional antennas with narrow beamwidths. This presents an advantage as it makes the technology less vulnerable to interference from neighboring users [176]. Whereas the 2.4 and 5 GHz bands offer limited opportunities in future due to interference from non-CBTC users, 60 GHz might serve as a promising alternative.

D. Ad-hoc based trackside network

Ad-hoc IEEE 802.11-based networks are generally considered inefficient. Since nodes must forward packets for other nodes, the capacity of such networks degrades severely as the size of the network grows. The movements of nodes further degrades capacity as a result of an increased number of routing messages [179], [180].

An ad-hoc network, formed as a chain (or string) of nodes, where each node forwards packets to its neighboring node, might present a suitable candidate for a CBTC trackside network. However, studies show that such a network only offers a fraction of the capacity achieved by a single-hop network [179], [180]. This is due to contention in IEEE 802.11-based networks where nodes sending more often starve the rest of the nodes. Additionally, a node in such a chain doesn’t only have to contend with its two immediate neighbors but at least 3 or 4 neighbors [181].

Studies show that the capacity of such networks drops to one-half with each hop, and to 1/7 as the number of nodes increases beyond 10 [180], [181].
Most of these problems, however, are mainly relevant in networks where nodes are transmitting on the same frequency. In [182], it is demonstrated that an IEEE 802.11-based ad-hoc network of forwarding nodes could serve as a promising alternative for CBTC scenarios. It proposes a three-frequency design, where a node forwards packets not only to its immediate neighbor but also to the next neighbor in the chain, to improve capacity and resiliency.

An advanced version of the ad-hoc based network is a Wireless Mesh Network (WMN). WMNs employ a multi-radio, multi-frequency design, and, despite their merits, have not been considered for CBTC, except in [30] and [96].

WMNs are dynamically self-organized, self-configured and self-healing networks, with the nodes in the network automatically forming an ad-hoc network, using multi-hop communication, to forward traffic to and from a wired network. A notable feature of these networks is their hierarchical architecture, where dedicated nodes, called "wireless routers", provide wireless connection from a user node to either other user nodes or APs [183], [184].

In conventional ad-hoc networks, mobility, particularly handover, has been a challenge. As a result of the multi-hop route discovery, the handover latency increases dramatically with the increasing number of hops [185]. However, the challenge might not be as prominent in CBTC because of the fixed nature of the CBTC wayside networks, where normally the train is the only mobile node.

IEEE 802.11s [186] is an extension to the IEEE 802.11 standard to support mesh networks. A wide range of vendors are currently offering competitive wireless mesh solutions. One example, targeted towards CBTC, is Fluidmesh [187], a Multiple Input and Multiple Output (MIMO) based product that claims to support 100 Mbps in metro tunnels. Additionally, a reference implementation of IEEE 802.11s is available in the Linux kernel starting from version 2.6.26 [188].

XI. Conclusions

Radio communication plays a key role in the modern communication-based signalling systems as it connects train to wayside to transfer high resolution and real-time train control information. With the help of this information, CBTC offers a number of major benefits over a conventional signalling system, namely, shorter headways resulting in greater capacity, fewer trackside equipment, greater punctuality, improved safety, and support for automated train operations. In this paper, we have aimed to present a comprehensive tutorial and a survey of the state-of-the-art of radio communication in CBTC. An overview of the evolution of communication technologies for railway signalling compared the radio-based communication to the early inductive loop-based and leaky waveguide-based communication. While these early technologies incur high installation and maintenance costs, the greatest challenge with the radio-based communication is interference. A thorough examination of the benefits and drawbacks of using a radio communication technology, in particular IEEE 802.11 Wi-Fi, for CBTC, showed that the success of Wi-Fi can mainly be attributed to its high data rates, ease of installation and maintenance, and its cost-effectiveness as a result of readily available COTS radio equipment and license-free operation. On the other hand, the susceptibility to interference, lack of support for mobility, and short signal range are some of its disadvantages. An overview of the fundamental components of a CBTC system, both onboard and wayside, as well as the three types of networks involved, has been presented. A comparison of the Wi-Fi based radio equipment to the early COTS and custom-built equipment proved the former to be of orders of magnitude cheaper. An in-depth overview of the CBTC radio network configuration showed that although a typical inter-AP distance in CBTC is only 200-300 meters because of the short range of Wi-Fi signals, significantly longer range of up to various kilometers could be achieved by adjusting various parameters. An evaluation of the alternative designs and topologies for the train-to-trackside radio network showed that redundancy is the key to providing high availability in CBTC, and the availability can be increased dramatically by ensuring redundancy at multiple levels. Roaming in a CBTC environment is an inevitable reality due to the short range of Wi-Fi networks and the high speeds of trains. Thus, a smooth handover from one Wi-Fi AP to another is a critical requirement. The IEEE 802.11 standard was primarily developed for stationary users within a limited area, and therefore inherently does not support mobility. A review of the roaming algorithm design for CBTC showed that a complex and intelligent roaming algorithm is thus a critical component of a CBTC radio communication system. A summary of different standardization efforts for CBTC has been included. With the intention to bring more attention to the IEEE CBTC standard, a brief summary of the standard has been presented, including the guideline parameter values for optimal performance. The discussion on the future research directions has highlighted a number of promising alternatives for the CBTC radio technology as well as the architecture of the CBTC radio network.

ACKNOWLEDGMENTS

The authors would like to thank Lars Bro, Kasper Tipsmark Therkildsen, Kennen Reinhard Iversen, Hans-Henrik Munch, Claus Winskov Jørgensen, Jens Elmo Andersen, Rasmus Thysstrup Karstensen, and Aleksander Sniady for their valuable feedback on the previous versions of this paper. Special thanks to Lars Bro for providing details on the antenna front-to-back ratio, the radio coverage study of the Copenhagen S-train CBTC system, and the hidden node problem in a CBTC scenario. Further thanks to Miakel Voss and Briand Cancellor for providing information on the Siemens study at Chongqing Rail Transit Line 1 in China, and on the previous versions of Siemens’ Airlink radio communication system, respectively. This work was mainly supported by Siemens and partly by Innovation Fund Denmark.

REFERENCES


[169] ——, “IEEE 802.11s draft:头-26b4a3f6eb606c21056e4f906a4dae88077346f5


Jahanzeb Farooq received his B.Sc. and M.Sc. degrees in computer science from Hamdard University, Pakistan, and Umeå University, Sweden, in 2004 and 2006, respectively. Jahanzeb has previously worked at the French Institute for Research in Computer Science and Automation (INRIA), France, where he was involved in the modeling and simulation of WiMAX networks. Since 2010, Jahanzeb has been working for Siemens A/S, Denmark, where he has been involved in the research and development of Siemens’ Communications-Based Train Control (CBTC) system. Recently, he has been involved in the implementation of the Copenhagen S-train CBTC project. Currently, he is pursuing his PhD—sponsored by Siemens A/S and partly by Innovation Fund Denmark—in collaboration with the Technical University of Denmark (DTU). Jahanzeb’s current research interests are related to design and performance enhancements for radio communication systems for train control.

José Soler received his M.Sc. degree in telecommunication engineering from the University of Zaragoza, Spain, in 1999 and his Ph.D. degree in electrical engineering from the Technical University of Denmark (DTU) in 2005. He has previously held positions with the Technological Institute of Aragon in Spain; the Electronics and Telecommunications Research Institute in South Korea; the Department of Communications, Optics, and Materials at DTU; and GoIP Aps in Denmark. He is currently an associate professor with the Networks Technology and Service Platforms Group of DTU Fotonik. His research interests are related to the integration of heterogeneous telecommunication networks and the application of telecommunication services into different domains.