Infocommunications Journal

A PUBLICATION OF THE SCIENTIFIC ASSOCIATION FOR INFOCOMMUNICATIONS (HTE)

June 2019 Volume XI Number 2 ISSN 2061-2079

Technically Co-Sponsored by IEEE ComSoc IEEE Communications Society

hte

IEEE HUNGARY SECTION

HTE for 70 years

Indexing information

Infocommunications Journal is covered by Inspec, Compendex and Scopus.

Infocommunications Journal is also included in the Thomson Reuters – Web of ScienceTM Core Collection, Emerging Sources Citation Index (ESCI)

Infocommunications Journal

Technically co-sponsored by IEEE Communications Society and IEEE Hungary Section

Supporters

FERENC VÁGUJHELYI – president, National Council for Telecommunications and Information Technology (NHIT)

GÁBOR MAGYAR – president, Scientic Association for Infocommunications (HTE)

Editorial Office (Subscription and Advertisements)**:**
Scientic Association for Infocommunications
H-1051 Budapest, Bajcsy-Zsilinszky str. 12, Room: 502
Phone: +36 1 353 1027
E-mail: info@hte.hu • Web: www.hte.hu

Articles can be sent also to the following address:
Budapest University of Technology and Economics
Department of Telecommunications and Media Informatics
Phone: +36 1 463 4189, Fax: +36 1 463 3108
E-mail: pvarga@tmit.bme.hu

Subscription rates for foreign subscribers: 4 issues 10.000 HUF + postage

Publisher: PÉTER NAGY

HU ISSN 2061-2079 • Layout: PLAZMA DS • Printed by: FOM Media

Impactful Surveys for the 70th Anniversary of HTE

Pal Varga

The impact of a scientific achievement is hard to measure, especially in the short run. Still, human lifespan is relatively short when compared to the wide spread applications of theoretical breakthroughs – so we define "factors" predicting the possible impact of new ideas and contributions.

Scientific surveys are considered a strange breed of articles, since it is not necessarily expected from them to present and validate novel ideas. They contribute to the fabric of human knowledge and recognition of the world around us in another way. Providing a pre-digested, comparative, comprehensive overview about the current achievements of a domain could be just as eye-opening for some, as a clear description of brand new findings.

This is the first year of a new decade for the Infocommunications Journal, and indeed, we can be proud of some impactful articles published in the first ten years. This year is remarkable for HTE, the Scientific Association for Infocommunications, our publisher, as well. HTE has been formed in the January 29, 1949, in Hungary – hence celebrating its 70th anniversary this year. The intention with this current issue is to start the new decade with some impactful overviews – surveys – and some breakthrough articles carrying novel ideas.

The seven papers of this issue includes three invited surveys and four papers that arrived to the open call. The invited papers cover three very current areas of the ICT domain: hacking of IoT (Internet of Things) devices, and surveys on Quantum Key Distribution (QKD) and Visible Light Communication (VLC). Papers from the open call are also targeting current interest: 5G networks, V2G (Vehicle to Grid) communications, IEEE 802.11ax, and advanced compression and indexing methods for massive data. Let us have a brief overview of these papers.

In their primer paper, Papp et.al. provide an introduction into hacking IoT devices. After introducing the basic background on the interfaces and the protocols at the hardware level, they summarize the methods and tools for extracting the firmware of the device and unpacking it for further analysis. Further, they give an overview on some basic firmware analysis methods and tools that can be used to find hard-coded passwords and keys, and to discover erroneous settings or bugs. Moreover, they describe some more advanced analysis methods that can be used to discover vulnerabilities in the binary programs that belong to the firmware.

Gyongyosi et.al. furnish a synopsis of the recent results of QKD. Their review focuses on the principles of discrete-variable and continuous-variable QKD (DVQKD and CVQKD) protocols, the main attributes of the recent implementations, as well as the integration of QKD into traditional and quantum communication networks.

Upon editorial request, Eszter Udvary created a comprehensive literature overview on Visible Light Communication, covering its features and applications. VLC has the potential to provide high-speed data communication with relatively good security and improved energy efficiency. After introducing the motivation for VLC technology development, the paper describes the main advantages and disadvantages of this technology, demonstrates the current challenges, discusses modulation techniques and finally, VLC applications.

Racz et.al. investigate the performance of the closed-loop control of an UR5 industrial robotic arm at varying network characteristics. They evaluated the differences of the intended and the realized trajectories of the arm, and correlated this with communication speed and latency. Further, they suggest a method to handle loss and jitter of robot control packets.

In their paper, Jako et.al. present a wireless authentication solution prototype, which allows electric vehicle owners to identify themselves nearby the charging station, but before connecting the plug to the Electric Vehicles. They built a conformance test system for the Supply Equipment Communication Controller in accordance with the ISO/IEC 15118 standards.

Islam and Kashem propose an OFDMA-based MAC protocol for IEEE 802.11ax named HTFA, which employs a hybrid mechanism for channel access. HTFA will provide high throughput of data as well as maintains improved fair access policy to the medium among the terminals.

A data structure is called (singly) opportunistic if it takes advantage of the redundancy in the input in order to store it in information-theoretically minimum space. Nagy et. al. propose R3D3 as a new tool for compressing and indexing bitvectors. R3D3 is, in contrast to previous work, doubly opportunistic, in that it realizes substantial space savings on the compressed data and the index alike.

Seventy years for our Association, and ten years for our Journal – this is a year for celebration: remembering some legendary achievements, and aiming for new challenges.

Pal Varga received the M.Sc. and Ph.D. degrees from the Budapest University of Technology and Economics, Hungary, in 1997 and 2011, respectively. He is currently an Associate Professor at the Budapest University of Technology and Economics. Besides, he is also the Director at AITIA International Inc. Earlier, he was working for Ericsson, Hungary, and Tecnomen, Ireland. His main research interests include communication systems, network performance measurements, root cause analysis, fault localisation, traffic classification, end-to-end QoS and SLA issues, as well as hardware acceleration. Recently he has been actively engaged with research related to Cyber-Physical Systems and Industrial Internet of Things. He has been involved in various industrial as well as European research and development projects in these topics. Besides being a member of HTE, he is a member of both the IEEE ComSoc (Communication Society) and IEEE IES (Industrial Electronics Society) communities, and the Editor-in-Chief of the Infocommunications Journal.

IoT Hacking – A Primer

Dorottya Papp, Kristóf Tamás, and Levente Buttyán

***Abstract*—The Internet of Things (IoT) enables many new and exciting applications, but it also creates a number of new risks related to information security. Several recent attacks on IoT devices and systems illustrate that they are notoriously insecure. It has also been shown that a major part of the attacks resulted in full adversarial control over IoT devices, and the reason for this is that IoT devices themselves are weakly protected and they often cannot resist even the most basic attacks. Penetration testing or ethical hacking of IoT devices can help discovering and fixing their vulnerabilities that, if exploited, can result in highly undesirable conditions, including damage of expensive physical equipment or even loss of human life. In this paper, we give a basic introduction into hacking IoT devices. We give an overview on the methods and tools for hardware hacking, firmware extraction and unpacking, and performing basic firmware analysis. We also provide a survey on recent research on more advanced firmware analysis methods, including static and dynamic analysis of binaries, taint analysis, fuzzing, and symbolic execution techniques. By giving an overview on both practical methods and readily available tools as well as current scientific research efforts, our work can be useful for both practitioners and academic researchers.**

***Index Terms*—IoT security, ethical hacking, penetration testing, embedded firmware analysis, binary program analysis.**

I. INTRODUCTION

THE Internet has grown beyond a network of laptops, PCs, and large servers: it also connects millions of small embedded devices. This new trend is called the Internet of Things, or IoT in short, and it enables many new and exciting applications. At the same time, however, it also creates a number of new risks related to information security.

On the one hand, embedding computers into everyday objects and connecting them to the Internet exposes our physical world to attacks originating from the cyber space. This means that cyber attacks may have physical consequences, including damage of physical equipment or even loss of human life. Probably, the most famous example for this is the Stuxnet worm [1], which was used in an attack targeting a uranium enrichment plant in Iran to compromise embedded industrial controllers and to physically damage the uranium centrifuges that they controlled [2]. Another famous example is the proof-of-concept attack on the Jeep Cherokee SUV [3], in which two security researchers remotely took control over a vehicle while it was running on the highway. Besides these famous cases, there are many other examples for cyber attacks on network connected embedded systems (essentially IoT applications), where the consequences were or could have been highly undesirable, including an attack on the Ukrainian power grid that resulted in an hour long black-out in the city of Kiev [4], an attack on a steel mill in Germany that resulted in "massive damage to the system" [5], and a potential attack that installed malware on pacemaker devices that could have resulted in a fatality [6].

The other side of the coin is that embedded devices with no or weak protection, when connected to the Internet, can put Internet based services and the Internet infrastructure itself at risk. Indeed, weakly protected WiFi routers, web cameras, and other "smart" devices connected to the Internet are low hanging fruits for attackers that they can use to build a massive attack infrastructure. An example for this is the Mirai botnet [7], which consists in millions of compromised IoT devices and which was used in the largest DDoS (Distributed Denial of Service) attack ever targeting the Domain Name System of the Internet and making popular Internet based services unavailable [8].

The general insecurity of the Internet of Things is a problem, and researchers have started to investigate what it stems from and how to address it. In a recent survey [9], the authors performed a comprehensive study on reported attacks and defenses in the IoT domain with the goal of understanding what goes wrong with existing IoT applications in terms of security. They identified 5 major problem areas: unconditional trust in the local network and in the physical environment an IoT device is operating in, over-privileging mobile applications used to control IoT devices, no or weak authentication, and implementation flaws. The study found that a major part of the attacks resulted in full adversarial control over IoT devices. The reason for this is that IoT devices themselves are weakly protected and they often cannot resist even the most basic attacks.

Whether IoT devices can be made more resistant to attacks in a cost efficient way is an open question and subject to intense research. However, even if future devices will be more secure, there are millions of devices already deployed, and it is also important to understand the level of security that they provide. This can usually be measured to some extent by penetration testing or ethical hacking methods. Hacking IoT devices can be fun, because it combines traditional hacking methods with some hands-on physical experience, but more importantly, it is also a very useful activity that can help discovering and fixing vulnerabilities in IoT devices that, if exploited, can result in highly undesirable conditions, as we saw above.

In this paper, we give a basic introduction into hacking IoT devices. We begin with giving an overview on hardware hacking, as IoT hacking is often started by disassembling the IoT device under study. The vulnerabilities that can be exploited to gain full adversarial control over a device can often be found in the device's firmware. Therefore, we con-

The authors are affiliated with the CrySyS Lab at the Department of Networked Systems and Services of the Budapest University of Technology and Economics, e-mail: (see http://www.crysys.hu/).
Kristóf Tamás is currently with Ukatemi Technologies.
Manuscript received: February 2019; revised: May 2019.

tinue our introduction by explaining how the firmware can be extracted from the devices and unpacked. Then, we briefly summarize some basic firmware analysis methods and tools that aim at identifying hard-coded secrets, misconfigurations of the device, and simple bugs in scripts. Most of these tools are open source and freely available on the Internet, and we provide references to them. Finally, we complete our primer on IoT hacking by providing a survey on more advanced analysis methods, including static and dynamic analysis of binaries, taint analysis, fuzzing, and symbolic execution. Advanced binary analysis of embedded firmware is still an active area of research, hence, instead of tools readily available on the Internet as in the case of basic firmware analysis, advanced methods are mainly described in scientific publications. Accordingly, we provide references to the most relevant papers in this exciting research domain. We hope that this duality (i.e., giving an overview both on practical methods and readily available tools, as well as on current scientific research efforts) makes our work useful for both practitioners and academic researchers.

II. HARDWARE HACKING

In the IoT context, the IoT device being analyzed is often physically accessible to the hacker, which allows him/her to inspect the hardware components of the device, including chips and connectors soldered on the motherboard, and peripherals attached to it. Inspection of the hardware can be carried out in three phases:

1) **Hardware reconnaissance without opening the device:** In this phase, the main objective is to collect publicly available information about the hardware at hand, mainly from the Internet, as whatever information is discovered in this phase can be used later in the analysis. For instance, the serial or model number printed on the device may allow for the identification of data sheets or manuals on the Internet, which might include important information about the device. Wireless devices produced or used in the USA have an FCC ID (Federal Communication Commission Identifier) printed on them, which one can use to look up information on different web sites[1]. These web pages usually contain more information about the device than its data sheet, including the labelled motherboard, I/O (Input/Output) pins, test reports, and external and internal photos about the device. For the later phases, it is vital to identify the power requirements of the device and the needed adapters. The most important information include the level of amperage, the level of voltage, and the polarity. From the data sheets and photos, or by visually examining the device, it is also important to identify whether it has any kind of tamper protection, because opening a tamper protected device can lead to irreversible damage of the hardware. Finally, it might be possible to obtain public information about some known vulnerabilities of the device, which may be exploited without opening the housing of the device.
2) **Opening the housing of the device and inspecting the motherboard:** This phase usually requires more electrical engineering knowledge. Most importantly, it might be impossible to re-assemble the device into its original state after dismantling. Therefore, photos and notes have to be made and taken during the dismantling process. Once the device is open, the chips, pins, and interfaces on it can be inspected. With the chip identifiers found, a search on different web databases[2] can determine the purpose of the chip (e.g. processor, flash, RAM) and the function of its pins. In addition, the external communication interfaces, such as UART (Universal Asynchronous Receiver-Transmitter) or JTAG (Joint Test Action Group), are identified in this phase, as well as signs of use of communication protocols, such as SPI (Serial Peripheral Interface) or I2C (Inter-Integrated Circuit).
3) **Desoldering the chips form the motherboard (if necessary):** Sometimes, the pins of a chip cannot be accessed without desoldering the chip from the motherboard. For instance, to dump the content of a flash chip, the chip might need to be desoldered from the motherboard in order to solder it to an external adapter with connectable pins.

At the end of this phase, profound knowledge is gained about how the analyzed IoT device works at the hardware level. The next stage could be dumping the firmware from the device via SPI, gaining root access to the device via UART, or looking for vulnerabilities using JTAG. We discuss these techniques in the following sections.

A. The UART interface and protocol

UART (Universal Asynchronous Receiver-Transmitter) is an asynchronous serial communication protocol. Being asynchronous, no external clock is required for synchronization, but communicating parties must agree on the speed of the communication, the so called baud rate. The most common baud rate values are 9600, 19200, 38400, 57600 and 115200 bps.

A hardware UART port has at least four pins: voltage (Vcc), Ground (Gnd), Transmit (Tx), and Receive (Rx). The Tx pin is used to transmit data from the device to another connected device, while the Rx pin is used to receive data from the other device. The communication is usually full duplex, meaning that both parties can transmit bits at the same time.

In IoT devices, the UART protocol is used to display debug information, or to configure or repair the device. For instance, if the device has a software malfunction and its web interface is unavailable, one approach to fix it is to make a wired connection to the device through its UART port. From the hacking point of view, UART can be used to collect information about the device's bootloader, operating system, and configuration. The steps to connect to an IoT device are the following:

[1] e.g., fccid.gov or fccid.io

[2] e.g., datasheets.com, arrow.com, datasheetcatalog.com, alldatasheet.com, microchip.com

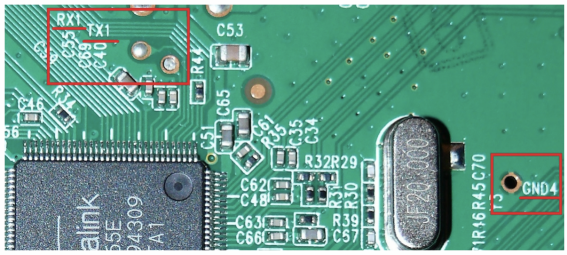


Fig. 1. UART ports on the TP-Link W8951ND router

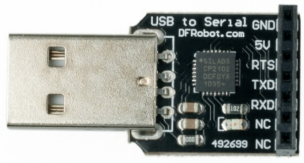


Fig. 2. An UART-to-USB converter device

- **Identifying UART ports and pinouts:** After removing the cover from the device, potential UART ports must be identified. The pins might be explicitly labeled on the motherboard or the four UART signals can be matched to the pins. In order to identify the pins, the board can be analyzed visually, or by using a multimeter or a logic analyzer. Figure 1 shows part of the motherboard of the TP-Link TD-W8951ND router where the UART pins are visible.
- **Connecting the UART pins to a computer:** After having identified the UART pins, the device has to be connected to a computer. For this step, special hardware is needed which can translate between USB and UART. These devices are usually called USB-to-TTL or UART-to-USB devices. An example is shown in Figure 2.
- **Identifying the baud rate:** In order to communicate with the device, the correct baud rate has to be identified.This can be done by trying the most common values manually. Also, there are open source scripts available for this purpose, such as `baudrate.py`[3].
- **Interacting with the device:** Besides the baud rate, the data frame configuration of the IoT device is also needed for proper communication. That can be determined in three ways: the vendor may have described it in the product manual, it might have been posted on a forum on the web, or it can be determined by trying the common frame configurations exhaustively. Once everything has been set, one can communicate with the device via UART by using off-the-shelf programs such as: `minicom`[4], `screen`[5], `dterm`[6], `picocom`[7], or `serialclient`[8]. To interact with a serial port, root privileges are required.

[3]https://github.com/devttys0/baudrate
[4]https://help.ubuntu.com/community/Minicom
[5]https://www.gnu.org/software/screen/manual/screen.html
[6]http://www.knossos.net.nz/resources/free-software/dterm/
[7]https://github.com/npat-efault/picocom
[8]https://github.com/flagos/serialclient

After a successful connection, some devices may require login credentials. Common username/password combinations can be tried to gain access to the device nevertheless. In other cases, UART connection to the device gives access to the boot-loader, a command line interface (CLI) or a shell. However, the received shell may be non-interactive, nevertheless, useful information can be gathered about the device.

B. The SPI protocol

SPI (Serial Peripheral Interface) is a synchronous serial communication bus protocol for short distance communication. SPI operates in a one-master-many-slaves setting, where one master (usually the CPU) controls a Slave Select (SS) wire for each slave. The master initiates communication with a slave by pulling down its SS wire. Also, the master is responsible for generating the clock signal. Like UART, SPI is also a full duplex protocol. Even when one party has no output to send, dummy data is sent on the affected line.

The communication takes place on four lines:

- Serial Clock (SCLK): The clock signal coming from the master. The clock speed must not exceed the maximum guaranteed clock speed of the selected slave.
- Master-Out-Slave-In (MOSI), sometimes Data In (DI): Communication line for sending data from the master to the selected slaves.
- Master-In-Slave-Out (MISO), sometimes Data Out (DO): Communication line for sending data from the selected slave to the master.
- Slave Select (SS): Signals to the slave that the master has initiated communication with it.

From the hacker's point of view, the SPI protocol is usually used to dump the content of an EEPROM or a Flash Memory, which typically implement the SPI protocol and store programs or persistent data.

Exploitation of SPI has similar steps to those of UART exploitation: Firstly, the chip of interest has to be identified and its pins must be matched to the lines described above. Then, the chip has to be connected to a computer, which can be done either with or without desoldering it. Communicating with the chip without desoldering is made possible by special clips such as the one shown in Figure 3. One challenge is that communication with the chip requires it to be powered up. Powering up the entire device is an option, but there can be interference on the chip's legs whenever the CPU tries to communicate with it. If the data sheet specifies the exact voltage level on which the chip should be used, the specified power can be directly applied to the chip from a DC power supply. If the legs are unreachable, then desoldering the chip and soldering it to an SPI Flash or EEPROM adapter is the only option left.

Communicating with the chip using the SPI protocol needs an SPI-to-USB adapter or bridge, such as Bus Pirate[9], a multifunction tool capable of UART, SPI, I2C and JTAG com-

[9]https://www.sparkfun.com/products/12942

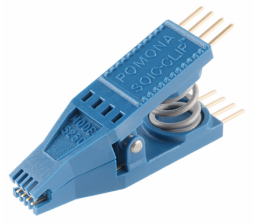


Fig. 3. SPI test clip

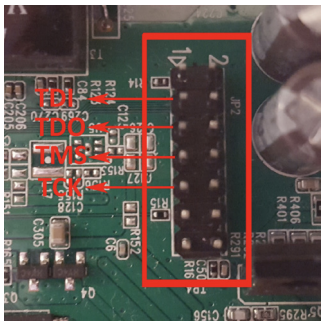


Fig. 4. JTAG interface

munications. Special software is also needed on the connected computer, such as `SPIFlash`[10], or `flashrom`[11].

C. The JTAG interface

JTAG (named after the Joint Test Action Group which specified it) is an industry standard for verifying designs and testing printed circuit boards after manufacture. Essentially, JTAG specifies the use of a dedicated port that implements a serial communications interface for accessing different signals on the board without requiring direct access to the system address and data buses.

JTAG has other higher level usage, namely debugging, which makes it possible to set breakpoints, view register and memory content, and dump the firmware. The beginning of the workflow to exploit a device through JTAG is quite similar to that of UART and SPI: identifying the JTAG pins (see Figure 4 for an example), connecting the device to a computer through JTAG and an adapter device like Bus Pirate mentioned above, and interacting with the device. Interaction is handled on the connected computer by an appropriate tool, such as OpenOCD[12].

OpenOCD uses special configuration files to communicate with the devices. There are many build-in configuration files, but new configurations can be created as well. However, this requires special knowledge about the device, such as its CPU architecture, endianness, TAP (Test Access Port) controller configuration, clock speed, *etc*. After finding or creating the configuration files, and connecting the device and the computer, OpenOCD accepts `telnet` connections at port 4444 and `gdb` connections at port 3333, which can be used to interact with the device.

[10]https://github.com/LowPowerLab/SPIFlash
[11]https://www.flashrom.org/Flashrom
[12]http://openocd.org

III. FIRMWARE EXTRACTION

The firmware is the low level code running on the IoT device that handles access to its hardware components and peripherals, and provides general services to higher level programs, such as an application. In this paper, we consider the operating system (if there is any) of the device as part of its firmware, which is a quite common approach in the domain of embedded systems.

The firmware usually consists of three main parts:

- **Bootloader:** A piece of low level code that initializes the hardware and loads the main operating system. Basically, it is the first program that is executed after switching a device on or after a reset. The bootloader might execute in two stages: in the first stage, only very basic code runs which loads code for the second stage, loading the operating system. This allows the second stage to be updated, while the first stage remains static. Common bootloaders used on embedded devices include Das U-Boot[13], MCU Boot[14], RedBoot[15], iBoot[16], BareBox[17], Bootbase and CFE[18]. Bootloaders may have vulnerabilities, which might be found by tools such as BootStomp[19], a bootloader bug finder for ARM architectures. Vulnerabilities in a bootloader may be exploited by malware, such as UbootKit [10], with the aim of loading a modified operating system and applications, i.e., to compromise the entire device.
- **Operating system (OS):** The operating system provides an execution environment for applications. The OS kernel is the core component of the operating system, which is loaded and started by the bootloader. There is a wide range of operating systems used in embedded devices, ranging from more complex ones like Linux to less complex ones like eCos. The most common operating systems used by IoT devices are Linux[20], VxWorks[21], eCos[22], OpenWRT[23], Junos OS[24] and uCOS[25]. Like the bootloader, the operating system might also contain security holes, but finding these are not trivial either. We discuss some of the approaches later in Section V.
- **File system:** The file system contains configuration files, libraries, development environments, and application programs run by the device. Many IoT devices ship with web servers on them, allowing for web based remote configuration of the device. Such applications are of particular

[13]https://www.denx.de/wiki/U-Boot
[14]https://github.com/runtimeco/mcuboot
[15]https://sourceware.org/redboot/
[16]https://www.theiphonewiki.com/wiki/IBoot_(Bootloader)
[17]https://www.barebox.org/
[18]https://en.wikipedia.org/wiki/Common_Firmware_Environment
[19]https://github.com/ucsb-seclab/BootStomp
[20]https://www.elinux.org/Main_Page
[21]https://www.windriver.com/products/vxworks/
[22]https://www.ecoscentric.com/ecos/index.shtml
[23]https://openwrt.org/
[24]https://www.juniper.net/us/en/products-services/nos/junos/
[25]https://www.micrium.com/rtos/

interest to hackers, because finding vulnerabilities in them does not require special embedded systems background. There are many different file systems for embedded devices including SquashFS[26], UBIFS[27], YAFFS2[28], and JFFS2[29].

A. Obtaining the Firmware

The firmware image of the IoT device can sometimes be found on the vendor's support page, although the image is often only partial. The complete firmware contains the entire file system, whereas a partial firmware image only contains some part of it (typically updated binaries or configuration files). However, even a partial firmware can reveal potential security flaws in older devices, because it usually contains updates that fix security holes. By comparing the updated files with their old versions, the vulnerability fixed in the update can be identified.

Even if the firmware image cannot be obtained from the vendor's support page, it may have already been made available on the Internet by other parties. However, firmware images obtained in this manner should be handled cautiously; they might be modified or their version might be different from the one on the device. The process is also time consuming, but in case of success, the exact binary that is present on the device can be obtained, potentially including the bootloader and the OS kernel. If the firmware image cannot be found on the Internet, it can be dumped from the device using the serial communication protocols presented in Section II.

Some devices have over-the-air (OTA) firmware update functionality, which can be initiated manually or automatically. During the update, a new (partial) firmware image is downloaded from the Internet, and hence, it can be captured with sniffing or man-in-the-middle techniques.

Finally, the simplest IoT devices like smart plugs, smart light bulbs, and smart locks usually come with mobile application used to manage them. Often such a mobile application contains a URL where the original firmware or firmware update can be downloaded from, but which is not indexed by search engines. Reverse engineering the mobile application can provide the hacker with that URL.

B. Unpacking the firmware

The complete firmware image is usually packed into a single compressed or archived file with the file system and the OS kernel. The file also contains a license file or user manual and a binary file. This binary file contains the firmware image, and is sometimes encrypted. The files packed within the binary may be further compressed or archived individually. In addition, the file system component can be stored in a special format. All in all, unpacking the firmware image usually requires to deal with encryption, compression and archive formats, and file system formats. The *de facto* standard tool used for unpacking is called `binwalk`[30], an advanced pattern matching tool capable of analyzing and extracting the content of a firmware image for a large number of different formats and encodings.

1) Dealing with encryption: Dealing with encryption is a challenge. The encryption algorithm and the entire encryption process might not be well-documented, and use proprietary methods. Even if a standard encryption algorithm, such as AES, is used, the keys are usually not readily available. The keys may be stored in tamper resistant hardware on the device, in which case, decrypting the firmware is near impossible. However, if the keys are stored in regular persistent memory which is not tamper resistant, then they can be extracted and the firmware can be decrypted.

To figure out whether the firmware is encrypted or not, entropy based analysis can be used, which is supported by `binwalk`. For an encrypted image, the entropy is flat across the entire binary and its value is close to 1. For a non-encrypted image, the entropy is not flat, its value is usually lower than 1, and it contains fluctuations across the entire file (i.e., there are sections with very low entropy values).

2) Dealing with compression: The different parts of the firmware are usually compressed or archived. Compression is used to save storage space, while archiving creates one single file from several files and directories. Compression and archives can be dealt with in almost the same way. There are many compression methods and archive formats, but `binwalk` can identify many of these methods and formats by searching for their *magic numbers* in the binary.

`binwalk` can find out if the file is compressed or archived even if the algorithm or format is unknown to the program, however, it cannot extract the content. In this case, one can try to find the decompression or extraction code in the non-volatile memory of the device. This, however, is difficult and requires deep technical skills.

3) Interpreting the file system: The file system becomes available after identifying, extracting, and decrypting the firmware image. It defines how files and directories are stored, accessed, and retrieved. A file system is just a binary blob in the firmware image, and its type can be identified based on signatures, just like in case of compressions and archives, but this method is typically more complex and less reliable for file systems. Extracting the file system content requires interpreting the structure, extracting the files, and placing them in the host file system. `binwalk` can identify and unpack many popular file systems, including those discussed at the beginning of this section.

4) What if `binwalk` fails?: It might seem for the reader that `binwalk` can unpack any firmware images. This is indeed true for common Linux-based firmware images in most of the cases. However, in case of special, proprietary firmware formats, `binwalk` may fail, as such formats may not use magic numbers or their extraction methods may be unknown to `binwalk`. However, even in such cases, `binwalk` may output useful information that can give clues regarding where to look for special tools that might work.

[26]http://squashfs.sourceforge.net
[27]http://www.linux-mtd.infradead.org/doc/ubifs.html
[28]https://yaffs.net
[29]http://www.linux-mtd.infradead.org/doc/jffs2.html
[30]http://binwalk.org

IV. BASIC FIRMWARE ANALYSIS

This section covers some basic analysis methods and tools, which can be used mainly on the non-binary parts of the firmware (e.g. text based config files and scripts in the file system). Analyzing the binary content requires advanced methods and tools, which we will discuss in details in Section V.

The main goal of basic firmware analysis is to find hard-coded secrets (e.g., passwords or keys) contained in non-binary files, such as configuration files, password files, and scripts (e.g., shell, Python, JavaScript, and Perl scripts, or alike). More specifically, the potentially collectable information include hard-coded credentials (e.g., username/password), private keys, encryption keys, API (Application Programming Interface) keys, access tokens, authentication cookies, and sensitive URLs or IP addresses. The file system may also store in readable format configuration files, lightweight database files, and password files that may contain useful information. In addition, it is also possible to figure out from the configuration files and scripts what services the device runs (e.g., `telnet`, `ssh`, `ftp`, `http`). Sometimes, basic firmware analysis may also include identifying common configuration errors in configuration files and exploitable programming bugs in scripts.

Useful tools for analyzing non-binary files include the following:

- **grep/egrep:** the *de facto* pattern matching tool on Linux. It can be used to search for strings like 'passwd', 'password', 'telnet', 'ssh', 'secret', *etc.* within all files in the file system.
- **find:** a tool that can find files by their attributes (content, name, permissions, type) with regular expressions.
- **firmwalker**[31]**:** a bash script that searches through the file systems for all the above mentioned keywords (passwords, keys, URLs, *etc.*) using `grep` and `find`, and saves the result in a text file.
- **firmflaws**[32]**:** a standalone Django web server, which uses other basic analysis tools to extract and analyze the contents of a firmware file. It expects a single packed firmware image as input, and it tries to extract its content (with `binwalk`) and analyze it.

A. Example: Basic analysis of the firmware of the D-Link DWR-932 WiFi router (version 4.00b05 Revision D)

For illustration purposes, we present here the result of our basic analysis of the firmware of the D-Link DWR-932 WiFi router.

We ran `firmwalker` on the firmware, which found nearly 1500 files. Some of those files contained interesting strings. We excluded the standard Linux binaries, and the HTML and JavaScript files, and manually analyzed the remaining files.

The file system contained the `/etc/passwd`, the `/etc/shadow` and the `/etc/group` files, which hold information about the users, their passwords, and the groups, respectively, on the system. Checking these files revealed that the default root password was empty:

```
$ cat /etc/shadow
root::17121:0:99999:7:::
...
```

Furthermore, the `/etc/securetty` file, which lists the terminals on which root is allowed to login, contained the serial console `ttyS0`, the USB dongle terminal `ttyUSB0`, and the standard consoles from `tty1` to `tty63`.

The file `/etc/miniupnpd/miniupnpd.conf` contains the default UPnP (Universal Plug and Play) configuration. UPnP was enabled on port 8201, with secure mode off and possible connections from any port and any host:

```
$ cat /etc/miniupnpd/miniupnpd.conf
...
port=8201
...
enable_upnp=yes
...
secure_mode=no
...
allow 0-65535 0.0.0.0/0 0-65535
```

We also found the possible WPA (WiFi Protected Access) passphrase 1234567890 and the possible WPS (WiFi Protected Setup) pin code 12345670 in multiple configuration files.

We identified that Dropbear (a lightweight SSH service) was present in the firmware, however, its automatic start was commented out:

```
$ cat /etc/init.d/dropbear
...
#start-stop-daemon -S \
#  -x "$DAEMON" -- $KEY_ARGS \
#    -p "$DROPBEAR_PORT" $DROPBEAR_EXTRA_ARGS
```

Dnsmasq 2.55 (a lightweight DNS and DHCP server) was also present and started automatically. However, versions lower than 2.78 have serious known vulnerabilities[33], although they can only be exploited when Dnsmasq is configured as a DHCPv6 server, which was not the case on this router.

V. ADVANCED FIRMWARE ANALYSIS

In this section, we give an overview on some advanced techniques used for uncovering vulnerabilities in firmware. As the presented techniques can focus on either the firmware image or the binary executables stored on the filesystem, we will refer to the analyzed piece of code simply as binary code.

Traditionally, analysis techniques can be categorized as either static or dynamic analysis techniques. *Static* techniques interpret instructions of the binary code and perform analysis in an abstract domain. These techniques scale well and can handle large code bases which makes them particularly useful for analyzing whole firmware images. Additionally, they require no test bed or platform. Coupled with the previous advantage, static analysis is a natural choice for performing large-scale analysis of firmware images [11]. However, without runtime information, such techniques often produce false positives, e.g. report vulnerable segments of code which cannot be executed in real life.

[31] https://github.com/craigz28/firmwalker

[32] https://github.com/Ganapati/firmflaws

[33] https://github.com/google/security-research-pocs/tree/master/vulnerabilities/dnsmasq

On the other hand, *dynamic* techniques analyze code as it runs on its intended platform. As a result, these techniques have access to runtime information, which allows for more precise results. However, dynamic techniques cannot provide information on behavior which has not been observed. As a result, these techniques are prone to false negatives, e.g. not all vulnerabilities may be reported. Additionally, analysis requires a test environment, which poses several challenges for IoT devices.

The advantages and disadvantages of both categories are complementary to each other and are often combined to achieve better results. Static analysis techniques are usually performed first, in order to focus dynamic analysis techniques to potentially vulnerable parts of the analyzed piece of code. In return, dynamic techniques can verify the results of static analysis and reduce false positives. As a result, the most advanced analysis techniques in literature cannot be categorized as either static or dynamic analysis, but instead inherit techniques from both categories.

The remainder of this section is structured as follows. We discuss the challenges of analyzing binary instructions in Section V-A and those of test environments for dynamic analysis techniques in Section V-B. Then, we discuss approaches to quickly find potentially vulnerable components in Section V-C and present three advanced analysis techniques: taint analysis in Section V-D, fuzzing in Section V-E and symbolic execution in Section V-F.

A. Challenges of analyzing binary instructions

In order to start analysis, the entry point of the binary has to be determined. This is easy for applications in known file formats (e.g. ELF for Linux-based systems), but challenging for proprietary formats and the firmware image itself. [12] overcame this challenge by analyzing jump tables in the image and starting analysis from multiple potential addresses.

In addition, precise analysis requires context sensitivity, i.e., all call and return sites have to be recovered accurately. While certain architectures have specific instructions for calling and returning from functions, other architectures can achieve the same semantics with indirect jumps. As an example, let us consider the ARM platform, in which the program counter (`pc`) is a general purpose register and the return address is stored in the link register (`lr`). The following (non-exhaustive) list of instructions all result in returns from functions:

```
; Push-pop pair
push lr
pop pc

; Unconditional jump
bx lr ; Used in functions where
      ; lr is not stored on the stack

; Direct program counter manipulation
mov pc, lr

; Bitwise operations
orr r15, r14, r14    ; pc (r15) = lr (r14)
                     ; bitwise-OR lr (r14)
```

While dynamic analysis tools have runtime information available and can accurately compute the call and return addresses, static analysis techniques are hampered in such scenarios. Additionally, there are proprietary architectures in the IoT ecosystem with unknown calling conventions, which makes streamlining tools a challenge [13].

The IoT ecosystem is a heterogeneous ecosystem with many architectures, platforms and firmware. This setting presents several challenges for interpreting binary code and performing static analysis. Firstly, compiler optimization heavily affects the resulting binary code and as a result, the same source code can be compiled into syntactically different, but semantically equivalent binary instructions. Secondly, the different architectures and calling conventions present in the IoT ecosystem make it hard to detect that two sets of instructions compute the same semantic result. Thirdly, depending of the toolchain used to compile a piece of code, the resulting binaries may differ as well.

To overcome these challenges and provide platform indepdence, static analysis techniques are typically not performed on the binary instructions but rather on an *intermediate representation* (IR). The instructions of an IR are often at a higher level than the binary instructions, however, they still lack the same semantic information found in source code. Popular intermediate representations include VEX of valgrind [14], TCG of QEMU [15] and the LLVM bitcode [16].

B. Setting up a test environment

Dynamic analysis techniques require an analysis environment in which the analyzed code can be run. If analysis has access to the underlying hardware or device, those could be used as the environment. However, most platforms do not ship with the tools required to turn the device into a test bed. As a result, significant engineering work is required before analysis can take place. If analysis has no access to the underlying hardware, there are multiple approaches to emulate the platform or certain parts.

Hardware emulation emulates all hardware elements of the underlying platform, including all its peripherals and interrupt handling system. This approach works well for well-understood platforms (e.g. QEMU [15]). However, the platform may be customized without accessible documentation which makes adapting existing emulators near infeasible. Vendors may develop accurate system emulators as part of the development lifecycle to enable firmware developers to work parallel to hardware developers. However, such emulators are usually unavailable to the public and often lack support for code instrumentation necessary for many security analysis techniques.

Even if the hardware is unavailable, the kernel could still be recovered from the firmware image. Emulating the recovered kernel can give more accurate analysis results. However, the kernel may be customized to the platform, hampering generic emulators. [17] overcame this limitation by leveraging the real device to handle I/O operations, signals and interrupts.

If the kernel cannot be recovered, the file system can still be booted with a generic kernel [18], assuming that the original

kernel is based on a generic, available kernel. Applications on the device can be analyzed, but analysis will not yield precise results if they rely on a customized kernel.

If analysis concerns only a single binary application from the file system and the kernel is not customized, then a generic environment can be emulated for the analyzed application, constraining its access to objects present on the original file system. In case of Linux-based IoT platforms, this approach relies on the Linux kernel's ability to call an interpreter to execute an ELF (Executable and Linkable Format) executable for a foreign architecture.

C. Finding potentially vulnerable components

Many vendors in the IoT ecosystem reuse open source components, e.g. the Linux kernel for firmware images, which are customized during the development process [19]. Security vulnerabilities in final products may come from the original code, as was the case with the Heartbleed vulnerability[34], or introduced to the product during development. Either way, the IoT ecosystem is left with potentially tens of thousands devices with similar vulnerabilities. Significant effort has been put into finding similar components based on known vulnerable functions. The main challenge in this area is the sheer number of devices and firmware images to cover.

In order to perform efficient searches, similarity metrics and bug patterns are required. Similarity metrics often include structural features [20], [21], [22], [23] such as the number of instructions, string and numeric constants or the structure of the control flow graph (CFG). However, such metrics face challenges when vulnerable components must be matched in a cross-platform manner. As discussed before, different platforms and toolchains can produce vastly different binary code, even if the original source code is the same.

To handle the heterogeneity of the ecosystem, similarity metrics capturing semantic information are required. Existing approaches include checking input-output pairs computed over higher-level representations, e.g. blocks of IR instructions [24] and conditional formulas [25]. Recently, machine learning algorithms have also been leveraged in order to quickly find code similar to a known vulnerable component [26], [27], [28].

D. Taint analysis

Taint analysis is a technique to detect vulnerabilities resulting from improper data sanitization, i.e. data derived from untrusted input is used in a security sensitive operation. The starting points of the analysis are called *sources* and denote program points where untrusted, user-controlled data can enter the analyzed piece of code, e.g. by reading environmental variables or reading from the standard input. The end points of the analysis are called *sinks* and denote security sensitive operations which can be utilized by attackers to carry out attacks, e.g. jump instructions for circumventing intended control flow. During analysis, the untrustworthiness of data is signaled by *tainting* it and then propagating the taint throughout the code according to a *taint propagation policy*. Vulnerabilities are detected, if a sink performs operations on tainted data. Note, however, that program integrity may have been violated before detection. Taint analysis has been successfully used to identify security-related crashes [29], [30] and recognizing protocol parser code in firmware images [31].

[34]https://www.wired.com/2014/04/heartbleed-embedded/ Last visited: 04.02.2019

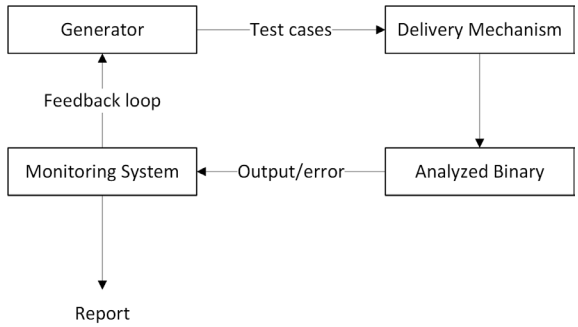


Fig. 5. Main Components of Fuzzing Tools

The technique can be performed in either static or dynamic ways. Static taint analysis [32] considers all possible execution paths starting from sources to sinks but faces the challenge of accurately identifying and analyzing data flows. Challenges arise from indirect memory accesses, indirect calls and pointer aliasing, when the same memory chunck is pointed to by different names. However, if the device cannot be emulated accurately, taint analysis can only be performed in a static manner.

Dynamic taint analysis [33], on the other hand, analyzes a single execution path and as a result, is able to handle scenarios challenging for static variants. However, certain challenges still remain. *Undertainting* is the error arising from the improper handling of certain information flows. Since taint analysis inherently deals with data paths, adding data dependencies to the taint propagation policy is obvious. However, information flow may occur through control dependencies as well, which cannot be computed in pure dynamic analysis as it requires considering multiple execution paths.

Overtainting (also known as *taint spread*), on the other hand, is the error of marking values tainted when they are derived from a taint source. For example, the ARM instruction `eor r0,r1,r1` computes the bitwise exclusive OR on `r1` and itself and then stores the result in `r0`. No matter the value of `r1`, the result will always be `0`. However, if the value of `r1` is tainted and the taint propagation policy does not exclude the example scenario, analysis will incorrectly consider the result tainted as well.

E. Fuzzing

The main idea behind fuzzing [34] is to supply randomly generated input values to the analyzed piece of code and then observe how it reacts. Since the input value is random, there is a high chance that it does not conform with the specification and will trigger anomalies in the code [35], [36], [37].

Figure 5 shows the high-level overview of the main components of fuzzing tools. The *generator* is tasked with generating

the random inputs used during analysis. There are three main types of strategies for input generation:

- **Mutation-based strategy:** Inputs are generated as a mutation of valid initial inputs. Initial inputs have to be specified at the beginning of the fuzzing process. This strategy is easy to set up even without a priori knowledge about the analyzed code, but has a low chance to pass validation checks.
- **Generation-based strategy:** Requires knowledge of program input, usually in the form of a configuration file. Generated random inputs confirm with the configuration file and are able to pass validation checks in programs, reaching deeper code.
- **Evolutionary strategy:** A feedback loop is used to supply the generator with information regarding execution behavior as well as results of other program analysis techniques. This allows for more fine-grained input generation and can greatly increase code coverage.

State-of-the-art fuzzing tools, like VUzzer [38], afl [39], or PULSAR [40], usually deploy evolutionary strategies.

The *delivery mechanism* receives the generated random inputs and supplies it to the analyzed binary. Depending on the input, different types of delivery mechanisms are needed, e.g. messages received over the network have to be delivered to the analyzed binary in a way that is different from the user behavior-based inputs on the embedded web server's graphical interface.

The *monitoring system* plays a crucial role in observing the output of the analyzed binary and detecting faulty behavior. In case of IoT devices, implementing a monitoring system is especially challenging because many traditional signals of faulty behavior are not present on these devices. What is more, the effects of memory corruption are often less visible because the analyzed piece of code may become unresponsive or produce late crashes [41]. There are two main approaches to implementing the monitoring system. Active probing requires special inputs to the code to check liveness. This approach was demonstrated in [42], where heartbeat messages were sent to the analyzed device over UDP. Passive probing, on the other hand, retrieves information about the execution state without alteration.

Fuzzing IoT devices presents unique challenges. In traditional IT settings, many instances of the same software can be started and fuzzed in parallel. For embedded devices, a large number of the same physical device is needed as many of them do not have the necessary memory and computational power, resulting in increased costs. Emulating the device could be a solution, however, there could be infrastructural limits to the number of devices that can be emulated in parallel. What is more, after a bug is triggered in the device, a clean state has to be restored, which often means a full reboot, slowing the process down.

Additionally, many tools require source code instrumentation to implement the feedback loop or the monitoring system. In the IoT ecosystem, the source code is often not available. Even when it is, a comprehensive toolchain would be required to recompile it into binary code. As a solution, dynamic binary instrumentation was proposed and is implemented in many tools, e.g. valgrind [14], Pin [43] or DynamoRIO [44].

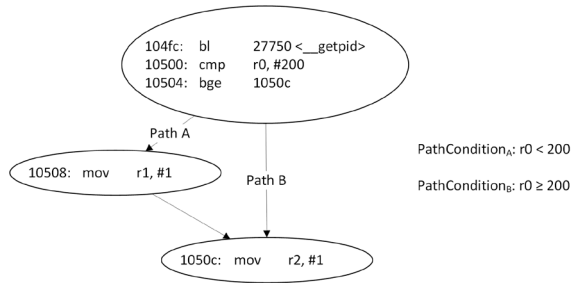


Fig. 6. Example ARM Instructions For Demonstrating Symbolic Execution

F. Symbolic execution

Symbolic execution is an emerging technique for finding vulnerabilities in IoT firmware images and applications [45]. This technique uses special symbols, *symbolic variables*, as values instead of concrete values to explore execution paths. Throughout this section, we demonstrate symbolic execution through the example ARM instructions in Figure 6. The code snippet calls `getpid()` and executes the instruction at 0x10508 only if the process ID is less than 200.

The first step of the analysis is the introduction of symbolic variables. Initially, symbolic variables are unconstrained representing the fact that a certain register or memory location may contain anything. In our example, consider the return value of `getpid()` an unconstrained symbolic variable.

The potential values of symbolic variables are refined at instruction which results in a control flow transitions. If multiple addresses can be followed, execution splits into multiple instances (*forks*). Each instance follows a potential control flow transition and places constraints upon the symbolic variables. The constraints represent the fact that the actual value held in the register or memory location had to satisfy the condition encoded into the branch. In our example, two execution paths are possible. On Path A, the constraint added to the path condition tells that the symbolic variable held in `r0` has to be less than 200. On Path B, the added constraint is for the symbolic variable to be greater or equal to 200. The constraints collected on an execution path are collectively referred to as the *path condition*. Note, that the path condition is taken into consideration at forks as previously added constraints may limit the available execution paths.

When an execution path terminates, the path condition can be solved by a Satisfiability Modulo Theory solver to acquire concrete values for the symbolic variables. The concrete values can then be used as test cases: assuming deterministic code, real-life execution of the analyzed piece of code will follow the same execution path as symbolic analysis did. In our example, for Path A, the solver could return any number below 200, e.g. 100. For Path B, it will return a value greater or equal to 200, e.g. 200. These concrete values can then be used to construct concrete test cases for both paths, maximizing code coverage automatically.

The concepts of symbolic analysis present multiple challenges for its real-life applications. As the subject has been discussed in multiple surveys [33], [46], we only give short descriptions of certain challenges as they are encountered during binary analysis.

Firstly, as analysis spawns two instances at each branch, the number of execution paths available for analysis grows exponentially, presenting serious *scalability issues*. There are many program constructs frequently used which result in exponential growth in the number of paths: symbolic loop guards, symbolic indices, etc. For binary code, symbolic offsets in memory and symbolic jump addresses can further complicate analysis. There two existing approaches to mitigate the issue:

- **Mixed concrete and symbolic execution**: The analyzed code is segmented into two parts: interesting instructions are analyzed over the symbolic domain, while uninteresting instructions are analyzed as if they were executed by the CPU. Segmentation can be determined by the tool: before an instruction is analyzed, the engine can check whether any of the operands is symbolic. If there is such an operand, the instruction is analyzed over the symbolic domain, otherwise it is analyzed over the concrete domain. However, given the complexity of some firmware images, the search space still remains too large for the mixed approach. In such cases, program slices can be computed over the firmware image to limit the scope of the analysis [12].
- **Path selection**: Unless the number of symbolic variables is kept at a minimum, the symbolic domain of the mixed approach may still remain too large to cover. As tools cannot hope to explore all execution paths, certain execution paths must be prioritized or abandoned according to some criterion. There have been numerous proposed criteria [47] for different application domains, but no universally effective method has been proposed yet.

Secondly, symbolic analysis engines have to model the execution environment of the analyzed code. In case of binary code, the engine has to posses knowledge about the potential registers a given platform can use, it has to model the memory and it must also be able to model the side effects certain instructions have (e.g. by setting flags) as well as interrupts [47] and hardware interactions [48]. In order to achieve platform independence, tools can leverage the intermediate representations discussed in Section V-A.

Finally, symbolic execution can only reason about code it analyzes, it cannot reason about unseen code. This challenge arises when specific binary applications are analyzed separated from the firmware image's filesystem and/or kernel and it is known as the *environment problem*. Unseen code (e.g. library functions, system calls) can have significant side effects on the analyzed program, which must be taken into consideration for precise analysis. One widely used solution is to create summary functions for such code to model its side effects. Several symbolic analysis tools (e.g. KLEE [49], EXE [50], angr [51]) implement this approach.

VI. CONCLUSION

In this paper, we gave a basic introduction into hacking IoT devices. We first introduced some details on the interfaces and the protocols at the hardware level that can be useful in a penetration testing context, and we explained how these interfaces can be identified in the device and how the protocols can be used for interacting with the device. Next, we summarized the methods and tools for extracting the firmware of the device and unpacking it for further analysis. We also gave an overview on some basic firmware analysis methods and tools that can be used to find hard-coded passwords and keys, to identify erroneous configuration settings, and to find simple bugs in scripts. Finally, we dealt with some more advanced analysis methods that can be used to discover vulnerabilities in the binary programs that belong to the firmware. Binary program analysis is still an active area of research, so we surveyed the most relevant scientific publications in the domain, including papers on static and dynamic analysis of binaries, taint analysis techniques, fuzzing, and symbolic execution of programs.

We deliberately restricted ourselves to hardware hacking and the analysis of the device's firmware, as vulnerabilities in the firmware can lead to full adversarial control over the device. We note, however, that penetration testing can be extended to the wireless interfaces of and protocols used by the device, to the applications running on the device, including web servers and remote access tools, to the mobile application that may be provided to remotely configure and control the device, and to the cloud end-points that the device may connect and send data to.

Ethical hacking of IoT devices is fun and useful at the same time. However, it is a relatively new area of research that still needs to mature. We expect that similarly to the best practice guides and standards for penetration testing of networks and web based applications, best practices and standards for IoT hacking will emerge and evolve in the near future. In particular, standards for security testing industrial IoT systems as well as autonomous and connected vehicles are needed in order to integrate the security testing activity into the development life cycle of those systems, and hence, to increase trust in them. In the home IoT area, best practices for security testing can encourage vendors to pay more attention to security, and ultimately, raise the bar for attackers to a level that is acceptable by home users.

ACKNOWLEDGEMENT

The work of Dorottya Papp and Levente Buttyán has been supported by the SETIT Project[35] (Security Enhancing Technologies for the Internet of Things).

REFERENCES

[1] N. Falliere, L. O'Murchu, and E. Chien, "W32.Stuxnet dossier," Symantec Technical Report version 1.4., 2011.
[2] R. Langner, "To kill a centrifuge – a technical analysis of what Stuxnet creators tried to achieve," online: http://www.langner.com/en/wp-content/uploads/2013/11/To-kill-a-centrifuge.pdf, 2013.

[35]Project no. 2018-1.2.1-NKP-2018-00004 has been implemented with the support provided from the National Research, Development and Innovation Fund of Hungary, financed under the 2018-1.2.1-NKP funding scheme.

[3] A. Greenberg, "Hackers remotely kill a Jeep on the highway – with me in it," *Wired*, July 2015.

[4] ——, "Crash Override: the malware that took down a power grid," *Wired*, June 2017.

[5] K. Zetter, "A cyberattack has caused confirmed physical damage for the second time ever," *Wired*, January 2015.

[6] L. H. Newman, "A new pacemaker hack puts malware directly on the device," *Wired*, August 2018.

[7] M. Antonakakis, T. April, M. Bailey, M. Bernhard, E. Bursztein, J. Cochran, Z. Durumeric, J. A. Halderman, L. Invernizzi, M. Kallitsis, D. Kumar, C. Lever, Z. Ma, J. Mason, D. Menscher, C. Seaman, N. Sullivan, K. Thomas, and Y. Zhou, "Understanding the Mirai botnet," in *Proceedings of the Usenix Security Symposium*, 2017.

[8] G. M. Graff, "How a dorm room minecraft scam brought down the internet," *Wired*, December 2017.

[9] N. Zhang, S. Demetriou, X. Mi, W. Diao, K. Yuan, P. Zong, F. Qian, X. Wang, K. Chen, Y. Tian, C. A. Gunter, K. Zhang, P. Tague, and Y.-H. Lin, "Understanding IoT security through the data crystal ball: Where we are now and where we are going to be," online: https://arxiv.org/pdf/1703.09809.pdf, March 2017.

[10] J. Yang, C. Geng, B. Wnag, Z. Liu, C. Li, J. Gao, G. Liu, and W. Yang, "UbootKit: a worm attack for the bootloader of IoT devices," in *BlackHat Asia Conference*, 2018.

[11] A. Costin, J. Zaddach, A. Francillon, D. Balzarotti, and S. Antipolis, "A large-scale analysis of the security of embedded firmwares." in *USENIX Security Symposium*, 2014, pp. 95–110.

[12] Y. Shoshitaishvili, R. Wang, C. Hauser, C. Kruegel, and G. Vigna, "Firmalice-automatic detection of authentication bypass vulnerabilities in binary firmware." in *Network and Distributed Systems Security Symposium (NDSS)*, 2015.

[13] M. C. Ang Cui and S. J. Stolfo, "When firmware modifications attack: A case study of embedded exploitation," in *Network and Distributed System Security Symposium (NDSS)*, 2013.

[14] N. Nethercote and J. Seward, "Valgrind: A framework for heavyweight dynamic binary instrumentation," in *Proceedings of the 28th ACM SIGPLAN Conference on Programming Language Design and Implementation*, ser. PLDI '07. New York, NY, USA: ACM, 2007, pp. 89–100. [Online]. Available: http://doi.acm.org/10.1145/1250734.1250746

[15] F. Bellard, "Qemu, a fast and portable dynamic translator," in *Proceedings of the Annual Conference on USENIX Annual Technical Conference*, ser. ATEC '05. Berkeley, CA, USA: USENIX Association, 2005, pp. 41–41. [Online]. Available: http://dl.acm.org/citation.cfm?id=1247360.1247401

[16] J. Zhao, S. Nagarakatte, M. M. Martin, and S. Zdancewic, "Formalizing the llvm intermediate representation for verified program transformations," in *Proceedings of the 39th Annual ACM SIGPLAN-SIGACT Symposium on Principles of Programming Languages*, ser. POPL '12. New York, NY, USA: ACM, 2012, pp. 427–440. [Online]. Available: http://doi.acm.org/10.1145/2103656.2103709

[17] J. Zaddach, L. Bruno, A. Francillon, D. Balzarotti *et al.*, "Avatar: A framework to support dynamic security analysis of embedded systems' firmwares." in *Network and Distributed Systems Security Symposium (NDSS)*, 2014.

[18] D. D. Chen, M. Woo, D. Brumley, and M. Egele, "Towards automated dynamic analysis for linux-based embedded firmware." in *Network and Distributed Systems Security Symposium (NDSS)*, 2016.

[19] M. Liu, Y. Zhang, J. Li, J. Shu, and D. Gu, "Security analysis of vendor customized code in firmware of embedded device," in *Security and Privacy in Communication Networks*, R. Deng, J. Weng, K. Ren, and V. Yegneswaran, Eds. Cham: Springer International Publishing, 2017, pp. 722–739.

[20] S. Eschweiler, K. Yakdan, and E. Gerhards-Padilla, "discovre: Efficient cross-architecture identification of bugs in binary code." in *Network and Distributed Systems Security Symposium (NDSS)*, 2016.

[21] Q. Feng, R. Zhou, C. Xu, Y. Cheng, B. Testa, and H. Yin, "Scalable graph-based bug search for firmware images," in *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, ser. CCS '16. New York, NY, USA: ACM, 2016, pp. 480–491. [Online]. Available: http://doi.acm.org/10.1145/2976749.2978370

[22] P. Shirani, L. Collard, B. L. Agba, B. Lebel, M. Debbabi, L. Wang, and A. Hanna, "Binarm: Scalable and efficient detection of vulnerabilities in firmware images of intelligent electronic devices," in *Detection of Intrusions and Malware, and Vulnerability Assessment*, C. Giuffrida, S. Bardin, and G. Blanc, Eds. Cham: Springer International Publishing, 2018, pp. 114–138.

[23] H. Lin, D. Zhao, L. Ran, M. Han, J. Tian, J. Xiang, X. Ma, and Y. Zhong, "Cvssa: Cross-architecture vulnerability search in firmware based on support vector machine and attributed control flow graph," in *2017 International Conference on Dependable Systems and Their Applications (DSA)*, Oct 2017, pp. 35–41.

[24] J. Pewny, B. Garmany, R. Gawlik, C. Rossow, and T. Holz, "Cross-architecture bug search in binary executables," in *2015 IEEE Symposium on Security and Privacy*, May 2015, pp. 709–724.

[25] Q. Feng, M. Wang, M. Zhang, R. Zhou, A. Henderson, and H. Yin, "Extracting conditional formulas for cross-platform bug search," in *Proceedings of the 2017 ACM on Asia Conference on Computer and Communications Security*, ser. ASIA CCS '17. New York, NY, USA: ACM, 2017, pp. 346–359. [Online]. Available: http://doi.acm.org/10.1145/3052973.3052995

[26] Y. Li, W. Xu, Y. Tang, X. Mi, and B. Wang, "Semhunt: Identifying vulnerability type with double validation in binary code." in *29th International Conference on Software Engineering and Knowledge Engineering (SEKE)*, 2017, pp. 491–494.

[27] J. Gao, X. Yang, Y. Fu, Y. Jiang, and J. Sun, "Vulseeker: A semantic learning based vulnerability seeker for cross-platform binary," in *Proceedings of the 33rd ACM/IEEE International Conference on Automated Software Engineering*, ser. ASE 2018. New York, NY, USA: ACM, 2018, pp. 896–899. [Online]. Available: http://doi.acm.org/10.1145/3238147.3240480

[28] J. Gao, X. Yang, Y. Fu, Y. Jiang, H. Shi, and J. Sun, "Vulseeker-pro: Enhanced semantic learning based binary vulnerability seeker with emulation," in *Proceedings of the 2018 26th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering*, ser. ESEC/FSE 2018. New York, NY, USA: ACM, 2018, pp. 803–808. [Online]. Available: http://doi.acm.org/10.1145/3236024.3275524

[29] K. Eom, J. Paik, S. Mok, H. Jeon, E. Cho, D. Kim, and J. Ryu, "Automated crash filtering for arm binary programs," in *2015 IEEE 39th Annual Computer Software and Applications Conference*, vol. 2, July 2015, pp. 478–483.

[30] H. Jeon, S. Mok, and E. Cho, "Automated crash filtering using inter-procedural static analysis for binary codes," in *2017 IEEE 41st Annual Computer Software and Applications Conference (COMPSAC)*, vol. 1, July 2017, pp. 614–623.

[31] Y. Zheng, K. Cheng, Z. Li, S. Pan, H. Zhu, and L. Sun, "A lightweight method for accelerating discovery of taint-style vulnerabilities in embedded systems," in *Information and Communications Security*, K.-Y. Lam, C.-H. Chi, and S. Qing, Eds. Cham: Springer International Publishing, 2016, pp. 27–36.

[32] K. Cheng, Q. Li, L. Wang, Q. Chen, Y. Zheng, L. Sun, and Z. Liang, "Dtaint: Detecting the taint-style vulnerability in embedded device firmware," in *2018 48th Annual IEEE/IFIP International Conference on Dependable Systems and Networks (DSN)*, June 2018, pp. 430–441.

[33] R. Baldoni, E. Coppa, D. C. D'Elia, C. Demetrescu, and I. Finocchi, "A survey of symbolic execution techniques," *ACM Comput. Surv.*, vol. 51, no. 3, 2018.

[34] J. Li, B. Zhao, and C. Zhang, "Fuzzing: a survey," *Cybersecurity*, vol. 1, no. 1, p. 6, Jun 2018. [Online]. Available: https://doi.org/10.1186/s42400-018-0002-y

[35] Z. Wang, Y. Zhang, and Q. Liu, "Rpfuzzer: A framework for discovering router protocols vulnerabilities based on fuzzing," *KSII Transactions on Internet and Information Systems*, vol. 7, no. 8, pp. 1989–2009, 2013.

[36] W. Frisby, B. Moench, B. Recht, and T. Ristenpart, "Security analysis of smartphone point-of-sale systems," in *Proceedings of the 6th USENIX Conference on Offensive Technologies*, ser. WOOT'12. Berkeley, CA, USA: USENIX Association, 2012, pp. 3–3. [Online]. Available: http://dl.acm.org/citation.cfm?id=2372399.2372403

[37] K. Koscher, A. Czeskis, F. Roesner, S. Patel, T. Kohno, S. Checkoway, D. McCoy, B. Kantor, D. Anderson, H. Shacham, and S. Savage, "Experimental security analysis of a modern automobile," in *2010 IEEE Symposium on Security and Privacy*, May 2010, pp. 447–462.

[38] S. Rawat, V. Jain, A. Kumar, L. Cojocar, C. Giuffrida, and H. Bos, "Vuzzer: Application-aware evolutionary fuzzing," in *Network and Distributed System Security Symposium (NDSS)*, 2017.

[39] M. Zalewski, "American fuzzy lop," http://lcamtuf.coredump.cx/afl/, last visited: Feb 7, 2019.

[40] H. Gascon, C. Wressnegger, F. Yamaguchi, D. Arp, and K. Rieck, "Pulsar: Stateful black-box fuzzing of proprietary network protocols," in *Security and Privacy in Communication Networks*, B. Thuraisingham, X. Wang, and V. Yegneswaran, Eds. Cham: Springer International Publishing, 2015, pp. 330–347.

[41] M. Muench, J. Stijohann, F. Kargl, A. Francillon, and D. Balzarotti, "What you corrupt is not what you crash: Challenges in fuzzing embedded devices," in *Network and Distributed System Security Symposium (NDSS)*, 2018.

[42] J. Chen, W. Diao, Q. Zhao, C. Zuo, Z. Lin, X. Wang, W. C. Lau, M. Sun, R. Yang, and K. Zhang, "Iotfuzzer: Discovering memory corruptions in iot through app-based fuzzing," in *Network and Distributed System Security Symposium (NDSS)*, 2018.

[43] C.-K. Luk, R. Cohn, R. Muth, H. Patil, A. Klauser, G. Lowney, S. Wallace, V. J. Reddi, and K. Hazelwood, "Pin: Building customized program analysis tools with dynamic instrumentation," in *Proceedings of the 2005 ACM SIGPLAN Conference on Programming Language Design and Implementation*, ser. PLDI '05. New York, NY, USA: ACM, 2005, pp. 190–200. [Online]. Available: http://doi.acm.org/10.1145/1065010.1065034

[44] D. Bruening, T. Garnett, and S. Amarasinghe, "An infrastructure for adaptive dynamic optimization," in *Proceedings of the International Symposium on Code Generation and Optimization: Feedback-directed and Runtime Optimization*, ser. CGO '03. Washington, DC, USA: IEEE Computer Society, 2003, pp. 265–275. [Online]. Available: http://dl.acm.org/citation.cfm?id=776261.776290

[45] I. Pustogarov, T. Ristenpart, and V. Shmatikov, "Using program analysis to synthesize sensor spoofing attacks," in *Proceedings of the 2017 ACM on Asia Conference on Computer and Communications Security*, ser. ASIA CCS '17. New York, NY, USA: ACM, 2017, pp. 757–770. [Online]. Available: http://doi.acm.org/10.1145/3052973.3053038

[46] E. J. Schwartz, T. Avgerinos, and D. Brumley, "All you ever wanted to know about dynamic taint analysis and forward symbolic execution (but might have been afraid to ask)," in *2010 IEEE Symposium on Security and Privacy*, May 2010, pp. 317–331.

[47] D. Davidson, B. Moench, S. Jha, and T. Ristenpart, "Fie on firmware: Finding vulnerabilities in embedded systems using symbolic execution," in *Proceedings of the 22Nd USENIX Conference on Security*, ser. SEC'13. Berkeley, CA, USA: USENIX Association, 2013, pp. 463–478. [Online]. Available: http://dl.acm.org/citation.cfm?id=2534766.2534806

[48] N. Corteggiani, G. Camurati, and A. Francillon, "Inception: system-wide security testing of real-world embedded systems software," in *27th {USENIX} Security Symposium ({USENIX} Security 18)*, 2018, pp. 309–326.

[49] C. Cadar, D. Dunbar, and D. Engler, "Klee: Unassisted and automatic generation of high-coverage tests for complex systems programs," in *Proceedings of the 8th USENIX Conference on Operating Systems Design and Implementation*, ser. OSDI'08. Berkeley, CA, USA: USENIX Association, 2008, pp. 209–224. [Online]. Available: http://dl.acm.org/citation.cfm?id=1855741.1855756

[50] C. Cadar, V. Ganesh, P. M. Pawlowski, D. L. Dill, and D. R. Engler, "Exe: Automatically generating inputs of death," *ACM Trans. Inf. Syst. Secur.*, vol. 12, no. 2, pp. 10:1–10:38, Dec. 2008. [Online]. Available: http://doi.acm.org/10.1145/1455518.1455522

[51] Y. Shoshitaishvili, R. Wang, C. Salls, N. Stephens, M. Polino, A. Dutcher, J. Grosen, S. Feng, C. Hauser, C. Kruegel, and G. Vigna, "Sok: (state of) the art of war: Offensive techniques in binary analysis," in *2016 IEEE Symposium on Security and Privacy (SP)*, May 2016, pp. 138–157.

Dorottya Papp received the MSc degree in Computer Science in 2016 from the Budapest University of Technology and Economics (BME) and she is currently pursuing the Ph.D. degree under the supervision of Dr. Levente Buttyán. She joined the Laboratory of Cryptography and System Security (CrySyS Lab) during her undergraduate studies at BME. Since 2015, she has also taken part in research projects at the Austrian Institute of Technology (AIT). Her research interests include security assurance, program analysis and security testing of binary programs. She is involved in research projects in the domain of embedded systems and the Internet of Things.

Kristóf Tamás received the B.Sc. and M.Sc. degrees (Diploma with Honours) in Computer Science from the Budapest University of Technology and Economics (BME) in 2017 and 2019, respectively. As a student, he has done research on identifying malicious web pages with static analysis and on penetration testing the Internet of Things under the supervision of Dr. Levente Buttyán. After graduation, he remained an associate member of the Laboratory of Cryptography and Systems Security (CrySyS Lab) and a member of the Lab's talent management program. He also has a leading role in the university's Capture-The-Flag (CTF) team, called c0r3dump. Kristóf Tamás currently works at Ukatemi Technologies Kft. - a CrySyS Lab spin-off company - as a junior security engineer and penetration tester.

Levente Buttyán received the M.Sc. degree in Computer Science from the Budapest University of Technology and Economics (BME) in 1995, and earned the Ph.D. degree from the Swiss Federal Institute of Technology - Lausanne (EPFL) in 2002. In 2003, he joined the Department of Networked Systems and Services at BME, where he currently holds a position as an Associate Professor and leads the Laboratory of Cryptography and Systems Security (CrySyS Lab). He has done research on the design and analysis of secure protocols and privacy enhancing mechanisms for wireless networked embedded systems (including wireless sensor networks, mesh networks, vehicular communications, and RFID systems). He was also involved in the analysis of some high profile targeted malware, such as Duqu, Flame, MiniDuke, and TeamSpy. His current research interest is in security of cyber-physical systems (including industrial automation and conrtrol systems, modern vehicles, cooperative intelligent transport systems, and the Internet of Things in general). Levente Buttyán played instrumental roles in various national and international research projects, published 150+ refereed journal articles and conference/workshop papers, and co-authored multiple books and patents. Besides research, he teaches courses on applied cryptography and IT security at BME and at the Aquincum Institute of Technology (AIT Budapest), and he leads a talent management program in IT security in the CrySyS Lab. He also co-founded multiple spin-off companies, notably Tresorit, Ukatemi Technologies, and Avatao.

A Survey on Quantum Key Distribution

Laszlo Gyongyosi, Laszlo Bacsardi, *Member, IEEE*, and Sandor Imre, *Senior Member, IEEE*

***Abstract*—Quantum key distribution (QKD) protocols represent an important practical application of quantum information theory. QKD schemes enable legal parties to establish unconditionally secret communication by exploiting the fundamental attributes of quantum mechanics. Here we present an overview of QKD protocols. We review the principles of QKD systems, the implementation basis, and the application of QKD protocols in the standard Internet and the quantum Internet.**

***Index Terms*—Quantum key distribution, quantum cryptography, security, networking.**

I. Introduction

Security and cryptography are crucial aspects of our everyday network communications. Since traditional networking methods are vulnerable to a variety of attacks, classical data encryption cannot provide unconditional security for legal parties [1]. QKD protocols [2]–[29] enable legal parties to share secret keys with unconditional security. In contrast to traditional cryptographic methods that rely on the computational complexity of mathematical functions, the security of QKD is based on physical laws. Whereas traditional cryptography is vulnerable to computational power [30], QKD systems are resistant against unlimited computational power. QKD can protect our security when quantum computers [31]–[36] become available.

The No-Cloning Theorem [37] is a consequence of the fundamentals of quantum mechanics, stating that it is impossible to make a perfect copy of a quantum system. In a QKD setting, it enables the parties to detect any eavesdropping activity, since the presence of an eavesdropper adds noise to the quantum transmission. The secret key between the sender (Alice) and receiver (Bob) is established over a quantum channel [29], which can be realized by an optical fiber [1], [6]–[22] or by a free-space optical channel [23]–[25], [38], [39].

QKD protocols can be classified into several different classes depending on the applied modulation, the encoding and decoding attributes, and the physical implementation of the quantum channel. Here we review QKD systems and the main attributes of the recent implementations.

This paper is organized as follows. In Section II, the fundamental principles of QKD protocols are discussed. In Section III, the implementation basis is studied. In Section IV, an outlook on quantum Internet scenarios is presented. Finally, Section V concludes the paper.

The research reported in this paper has been supported by the National Research, Development and Innovation Fund (TUDFO/51757/2019-ITM, Thematic Excellence Program). This work was partially supported by the National Research Development and Innovation Office of Hungary (Project No. 2017-1.2.1-NKP-2017-00001), and in part by the BME Artificial Intelligence FIKP grant of EMMI (BME FIKP-MI/SC).

The authors are with the Department of Networked Systems and Services, Budapest University of Technology and Economics, 1117 Budapest, Hungary (e-mail: gyongyosi@hit.bme.hu, bacsardi@hit.bme.hu, imre@hit.bme.hu).

II. Quantum Key Distribution

The first QKD protocols that were introduced were based on discrete variables (DV), such as photon polarization. These QKD protocols are termed DVQKD systems [1]–[8], [10]–[21]. The first DVQKD protocol that was introduced was the so-called BB84 protocol [2], which used single-photon polarization for the encoding. In the BB84 protocol, the classical random bits are encoded in single-photon polarization photons (qubits) with four random polarization states. The four polarization states belong to two bases: the rectilinear basis and the diagonal basis. In the encoding and decoding phases, these bases are randomly selected to prepare and to measure the photons. After the quantum-level transmission is closed, the parties use a classical authenticated channel (public channel) to compare the bases. In a phase called the basis agreement phase, the parties delete those bits from the key that have different bases. After this step, additional calculations and error-correcting operations are performed on the classical bit string to reduce the possibility that valuable information is leaked to an eavesdropper. This step is the distillation phase. The result of this phase is an absolute secure key between Alice and Bob. A simplified version of the BB84 protocol is the B92 protocol [40], which uses only two polarization states instead of four.

In an entanglement-based QKD protocol, entangled photon pairs are shared between Alice and Bob to generate a secret key [3]. The effectiveness of this protocol can be improved by the application of hyper-entangled states [41] (photon pairs that are entangled simultaneously in multiple degrees of freedom), which can increase the eavesdropping detection probability. QKD protocols motivated the development of other quantum cryptographic protocols in which the primary aim is not the establishment of a secret key, such as quantum dense coding [42], quantum teleportation [43]–[46], quantum secret sharing [47], [48], or quantum-secured blockchain [49].

Since the polarization of single photons cannot be encoded and decoded efficiently because of the technological limitations of current physical devices, continuous-variable (CV) QKD systems were proposed [22], [50]–[63]. In a CVQKD system, the information is encoded in continuous variables (i.e., photon packets) by a Gaussian modulation utilizing the position or momentum quadratures of coherent quantum states. In comparison with DVQKD, the modulation and decoding of continuous variables does not require specialized devices and can be implemented efficiently by standard telecommunication networks and devices that are currently available and in widespread use. As a convenient consequence, CVQKD systems can be integrated into the currently established telecommunication networks by using the present optical fiber networks and optical devices. CVQKD protocols

can be further classified into one-way and two-way systems. In a one-way CVQKD system, Alice transmits her continuous variables to Bob over a quantum channel [29], [62], [63]. In a two-way system, Bob starts the communication, Alice adds her internal secret to the received message, and this is then sent back to Bob (e.g., one mode of the coupled beam that is outputted by a beam splitter is transmitted back to Bob). Two-way CVQKD systems were introduced for practical reasons to overcome the limitations of one-way CVQKD systems, such as low key rates and short communication distances [52]. Two-way CVQKD protocols exploit the benefits of multiple uses of the quantum channel and can leak only less valuable information to the eavesdropper.

The two-field (TF) QKD system [17] is a novel QKD scheme that uses a continuous-wave (CW) laser. In a TF-QKD system, pairs of phase-randomized optical fields are generated at two distant locations, which are then combined at a central measuring station. The fields that convey the same random phase can be used to establish a secret key.

We note that there are several other types of QKD protocols that are not detailed in our paper (such as coherent one-way (COW) QKD [64], differential phase-shift (DPS) QKD [65], six-state QKD [66], and decoy-state QKD systems [7], [67]).

A. Discrete Variable Quantum Key Distribution

1) Modulation: In a DVQKD system, the quantum signal source is a single-photon source (e.g., attenuated laser pulses with telecom wavelengths). In the modulation phase, Alice draws a uniform random bit string that constitutes her raw data, and she then encodes the bits of the raw data into single-polarization photons with four (in BB84 [2]) random polarization states that represent the qubits. In the BB84, these polarization states are $\{\rightarrow, \uparrow, \nearrow, \nwarrow\}$, i.e., the horizontal, vertical, diagonal right, and diagonal left states that encode the logical bits $\{0, 1\}$ in the $B_r = \{\rightarrow, \uparrow\}$ rectilinear and in the $B_d = \{\nearrow, \nwarrow\}$ diagonal basis, respectively. The qubits are therefore modulated via a B random basis selection procedure.

2) Eavesdropping: The activity of an eavesdropper (Eve) results in detectable noise in the quantum channel, since Eve has no knowledge about the basis of Alice's qubit. As a corollary, for some qubits she will use the same basis as Alice, while for others a different basis is used, which results in detectable noise. The resulting noise of Eve's activity is analogous to a binary symmetric channel (BSC), which allows the use of the well-known channel-coding and error-correction tools in the post-processing phase.

3) Measurement: In a DVQKD system, the single-polarization photons are measured in the B_d basis or in the B_r basis in a B' random basis selection procedure at the receiver. In BB84, Bob randomly uses a rectilinear or diagonal basis, and the result of the measurement is a logical bit. These measurement results comprise Bob's raw data. Since Bob has no knowledge about the correct basis for the measurement of a given photon, several bits from his raw data will be uncorrelated with Alice's raw data. These bits are deleted from the raw data in the basis agreement phase, which uses the classical public channel.

4) Key Distillation: Key-distillation is a post-processing step that is separated from the transmission of quantum states. It aims to derive the secret key from the correlated raw data at the parties. The logical layer-based post-processing consists of two main phases: error correction and privacy amplification. The aim of the post-processing is to extract as much valuable information from the correlated raw data as possible and to generate an error-free key between Alice and Bob. The privacy amplification operates on the shared, error-corrected common secret to extract the final key between the parties, and the aim of this phase is to reduce to zero the possible knowledge of an eavesdropper from the elements of the key. The raw data shared over the quantum channel is noisy, and this must be corrected to distill the final secret key. Since a large number of raw data bits must be shared between the parties, the complexity of the post-processing phase is a critical point in QKD protocols.

B. Continuous Variable Quantum Key Distribution

1) Modulation: A Gaussian modulation is a robust and easily applicable solution in a practical CVQKD scenario [62], [63]. In particular, Alice draws a random Gaussian vector (Alice's raw data) and encodes the position and momentum quadratures based on it. The quantum signal source is a multi-photon source (e.g., a laser source with telecom wavelengths). In the standard CVQKD coding scenario, Alice modulates and separately transmits a CV coherent quantum state in the phase space. This standard modulation scheme is referred to as single-carrier modulation throughout the paper, consistent with its traditional meaning. In a multicarrier CVQKD [38], [68]–[73], the information is granulated into subcarrier continuous variables in the encoding phase, which are then decoded by a continuous unitary transformation. The aim of multicarrier CVQKD is to improve the secret key rates and the achievable distances.

2) Eavesdropping: For any CVQKD protocol, the optimal attack results in Gaussian noise; therefore, the physical link is modeled as an additive white Gaussian noise (AWGN) channel (Gaussian channel). More precisely, the Gaussian noise of the quantum channel models the eavesdropper's optimal entangling-cloner attack, and the channel is referred to as a Gaussian quantum channel. CVQKD schemes use continuous-variable Gaussian modulation, which has been proven to provide optimal key rates against collective attacks at finite-size block lengths, in addition to maximizing the mutual information between Alice and Bob [22], [74]. The security of CVQKD has also been proven against collective attacks in the asymptotic regime with infinite block sizes [62], [63], [75] and against arbitrary attacks in the finite-size regime [62], [63], [76]. Compared with a DVQKD system, a CVQKD system requires several additional physical parameters (transmittance, variance, shot noise, excess noise, the variance of Eve's quantum state, etc.) for the proper description of a Gaussian quantum channel. The performance of the protocol is strongly determined by the excess noise of the quantum channel and the transmittance parameter of the physical link.

3) Measurement: The measurement phase is a crucial part of CVQKD protocols. Depending on the measured quadrature types, it can be classified as homodyne or heterodyne measurement [62], [63]. In a homodyne measurement M_{hom}, only one quadrature, the position or the momentum quadrature x_j of a j-th coherent state, is measured. In a heterodyne measurement M_{het}, both the position and momentum quadratures are measured. Each quadrature measurement results in a unit in the raw data. Bob's resulting raw data are in the form of a noisy Gaussian vector with additive Gaussian noise. The raw data themselves do not comprise a secret key; they consist only of the results of the random quadrature measurements. The secret key is a uniformly distributed long binary string, which will be combined with the raw data elements in the stage of logical layer manipulations. The post-processing phase uses a classical-authenticated communication channel and classical error-correction algorithms.

4) Reconciliation: The reconciliation process of correlated Gaussian variables is a complex problem that requires either tomography in the physical layer, which is intractable in a practical scenario, or high-cost calculations in the multi-dimensional spherical space with strict dimensional limitations. In the reconciliation phase, only uniform distributions can be transmitted over the classical channel; otherwise, the information-theoretic security of the protocol cannot be proven [62], [63]. The raw data follow a Gaussian random distribution because the data arise from a Gaussian random source; however, by applying some trivial operations on the raw data units, the desired uniform distribution can be reached, and the reconciliation can be performed with unconditional security [77]. In the reconciliation phase, a physical-logical channel conversion is made, and the aim is to get a logical channel (reconciliation channel) that is close to a binary Gaussian channel. At low signal-to-noise ratios (SNRs), the capacities of the Gaussian quantum channel and the binary Gaussian channel are close, and the reconciliation channel is analogous to a binary Gaussian channel. The efficiency of the channel conversion procedure can be described by the relevant parameters of the resulting logical binary channel (such as its variance and capacity). This conversion efficiency determines the efficiency of the reconciliation process, i.e., the performance of the protocol.

In Fig. 1, the DVQKD and CVQKD settings are compared. The modulation phase in the DVQKD setting assumes four polarization states of the BB84.

III. QKD IMPLEMENTATIONS

A. QKD over Optical Fiber

The optical fiber infrastructure provides a base ground for the experimental realization of both DVQKD and CVQKD protocols. The currently established optical fiber infrastructure with wavelength division multiplexing (WDM) technique represents an adequate solution for the practical implementation of QKD [8]. A general architecture of a QKD-integrated optical network consists of four layers: a physical layer with the optical fiber architecture (e.g., an optical layer), a QKD layer, a control layer (which can be implemented by software-defined networking, or SDN, to efficiently manage the entire network [6]), and an application layer. In the layer model, the users' service requests are generated in the application layer. Then, the control layer determines a path in the physical network and performs a handshake with the relevant quantum devices and optical nodes through the path. In an abstract manner, the optical layer integrates optical nodes connected by optical fibers, while the QKD layer consists of quantum nodes with quantum channels and public channels between them. The optical layer and the QKD layer share the fiber bandwidth resources with WDM technique [6], [8]. On the problem of wavelength allocation and channel isolation for QKD-integrated optical networks, we refer to [13]. For the model of SDN-controlled optical networks with time-shared QKD, see [15]. On the problem of efficient secret-key allocation in QKD implementations, we suggest [16]. In [20], a method for the implementation of quantum and classical signals over the same optical fiber in QKD networks has been proposed. In [21], the concept of a virtual optical network (VON) is defined for the purpose of efficient energy utilization and security enhancements in practical optical fiber settings.

B. Free-Space Optical QKD

The fundamental characteristics of optical fiber-based QKD (i.e., channel loss of fibers, propagation losses) limit the achievable point-to-point distances to a few hundred kilometers. The achievable distances in terrestrial free-space-based QKD are also limited because of the exponentially decreasing photon rate with increasing distance. Satellite-based QKD represents a way to overcome these drawbacks and to establish a global-scale QKD network [23]–[25], [38]. The satellite-based solutions exploit the negligible photon loss and decoherence in the empty outer space. In [39], a satellite-to-ground QKD system with an achievable distance of over 1,200 kilometers has been demonstrated. The proposed model integrated a low-Earth-orbit satellite with decoy-state QKD. The reported key rate of the protocol was above 1 kbps. The results also enable us to realize high-efficiency long-distance QKD in a global-scale setting.

Relevant attributes of some recent QKD implementations are summarized in Table I.

C. QKD in the Traditional Internet

The secret key generated by a QKD system is a random key that can also serve as a one-time pad (OTP) [78], which theoretically provides unconditional security [79]. However, in theory, in an OTP system, the secret-key size must be at least as long as the data size to be encrypted, and novel random keys are required for novel data. It is trivially not implementable in practical scenarios because of the long execution times and large storage requirements. These issues are resolved by the integration of QKD into efficient traditional data encryption algorithms (AES, IPSec, TLS, etc.) [12], [80]. In these integrated, hybrid QKD-traditional encryption systems, the QKD structure provides a practical and significantly shorter key (in comparison with an OTP key) to an efficient encryption method that periodically requires a novel key from the QKD backbone structure [6].

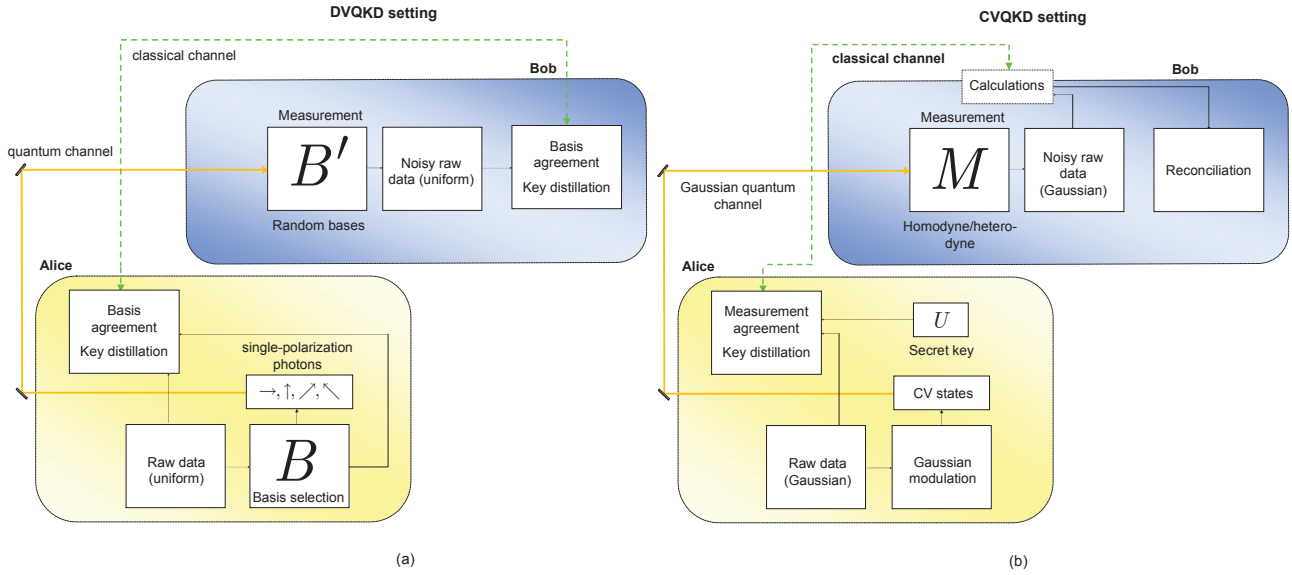


Fig. 1. Comparison of the sender (Alice) and receiver (Bob) model in a DVQKD and a CVQKD setting. (a) DVQKD setting. Alice draws uniform random raw data, which encode her random bits. She modulates all the bits of her raw data into single-polarization photons (qubits). The rectilinear and diagonal polarization states are selected randomly in the B basis selection procedure for the encoding. The qubits are sent through the quantum channel (depicted by the yellow line), where the presence of Eve adds noise to the transmission. Bob measures each qubit in a random basis via the B' basis selection procedure. The results of the measurements are classical bits, which form the noisy raw data. The final key is extracted from the correlated raw data of the parties using the classical public channel (depicted by the green line). (b) CVQKD setting. Alice draws Gaussian random raw data with Gaussian variables. Using her raw data, she modulates the CV quantum states via a Gaussian modulation. The CV quantum states are sent through a quantum channel, where the presence of the eavesdropper adds white Gaussian noise to the transmission. Bob measures the CV states via the M measurement procedure using homodyne or heterodyne measurement. The measurements yield noisy Gaussian raw data. In the post-processing phase, a U secret key (a classical uniform random vector) is drawn at Alice, which will be combined with her raw data. The combined result is transmitted to Bob over the classical channel. Bob applies some local calculations and reconciliation steps to extract the noise-free U secret key on his side.

TABLE I
ATTRIBUTES OF RECENT QKD IMPLEMENTATIONS.

QKD protocol	Distance	Max. secret-key rate	Quantum channel
BB84 (DV) [8]	66 km	5.1 kbps	optical fiber, 1310 nm
BB84 (DV) [10]	150 km	1 kbps	optical fiber, 1548 nm
BB84 (DV) [11]	80 km	1 kbps	optical fiber, 1310 nm
BB84 (DV) [18]	50 km	1.26 Mbps	optical fiber, 1550 nm
BB84 (DV) [19]	404 km	1.16 bit/hour	optical fiber, 1550 nm
Twin-field QKD [17]	550 km	0.1 kbps	optical fiber, 1550 nm
CV [9]	20 km	90 kbps	optical fiber, 1550 nm
CV [22]	80 km	0.1 kbps	optical fiber, 1550 nm
Satellite-to-ground BB84 (DV) [39]	1,200 km	1 kbps	free space optical, 850 nm

The hybrid structure is realizable through the currently established Internet architecture, as depicted in Fig. 2. The QKD devices establish the unconditionally secure key through the quantum channels (auxiliary public channels are not depicted). The keys are then passed via secure local connections to the server (e.g., an HTTP/TLS server) and the web clients. Then, the client-server communication is realized by the TLS protocol with periodically updated quantum-made keys.

IV. QKD IN THE QUANTUM INTERNET

The quantum Internet [80], [82]–[85] is a global-scale quantum communication network composed of quantum subnetworks and quantum networking components. The quantum Internet utilizes the fundamental concepts of quantum mechanics for networking. The main attributes of the quantum Internet are unconditional security (quantum cryptographic protocols), advanced quantum phenomena and protocols (such as quantum superposition, quantum entanglement, quantum teleportation and quantum coding and an entangled network structure. In contrast to traditional repeaters, quantum repeaters cannot apply the "receive-copy-retransmit" mechanism, because of the No-Cloning Theorem [37]. This fundamental difference between the nature of classical and quantum information not just leads to fundamentally different networking mechanisms, but also requires the definition of novel networking services in a quantum Internet scenario [86]–[90].

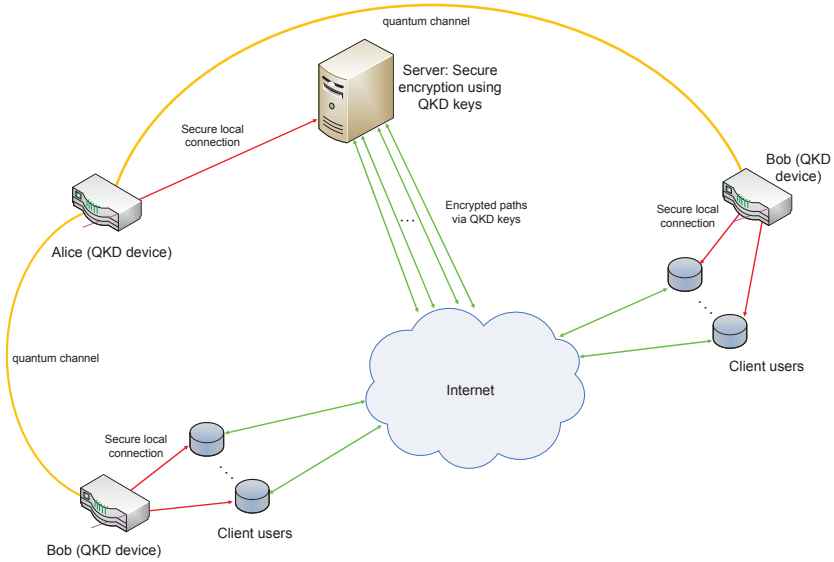


Fig. 2. QKD in a traditional client-server Internet setting. The established paths (green lines) between the clients and the server in the traditional Internet use quantum-made keys. The quantum keys are established via the QKD devices over quantum channels (depicted by yellow lines). The quantum keys are shared with the classical server and the classical clients through secure local connections (red lines).

The core network of the quantum Internet is modeled as an entangled network structure [80], [91], [92], in which the quantum nodes are connected by entangled connections. An entangled connection refers to a shared entangled system (i.e., a Bell state for qubit systems to connect two quantum nodes) between the quantum nodes. In an unentangled network structure, the quantum nodes are not necessarily connected by entanglement [93], [94], and the communication between the nodes is realized in a point-to-point setting. This setting does not allow quantum communication over arbitrary distances, and an unentangled network structure can mostly be used for establishing a point-to-point QKD between the quantum nodes. These short distances can be extended to longer distances by the utilization of free-space quantum channels [23], [24], [80], [95]. However, this solution is auxiliary, since it can be used only at some specific points of the unentangled network structure. Therefore, it does not represent an adequate and fundamental answer to the problem of long-distance quantum communication. Consequently, in an unentangled network structure, the multi-hop settings are weak for experimental, long-distance and global-scale quantum communication. On the other hand, the entangled network structure allows the parties to establish multi-hop entanglement, multi-hop QKD, high-precision sensor networks, advanced distributed computations and cryptographic functions, advanced quantum protocols, and, more importantly, the distribution of quantum entanglement over arbitrary (unlimited, in theory) distances [80]. Entanglement between a distant source and a target node is established through several intermediate repeater nodes [80], [91], [92], [96], [97]. The level of entanglement (i.e., the level of an entangled connection) is defined as the number of nodes (i.e., the hop-distance between entangled nodes) spanned by the shared entanglement, whose range is extended by the basic operation of entanglement swapping (entanglement extension).

The entangled network structure of the quantum Internet formulates a high-complexity network space with several advantages and challenges. Quantum Internet is an adequate answer for the computational power that became available as quantum computers became publicly available. The structure of the quantum Internet keeps the data of users safe for future networking. However, the commercial quantum computers are currently under development and represent tomorrow's problems, the engineering of high-performance and well-designed services and protocols for the quantum Internet is today's tasks. As quantum computers are built and become available, the structure of the quantum Internet also has to be ready to provide a seamless transition from the traditional Internet to the quantum Internet.

A. Recent Implementations

An optical switcher-based QKD implementation has been proposed in [14]. The system model integrates several hop-by-hop QKD settings to realize a long-distance QKD. The optical switchers were implemented for the purpose of time division multiplexing (TDM) on the quantum channels between the QKD devices.

A technical roadmap on the experimental development of the quantum Internet has been provided in [98]. The roadmap is connected to the Quantum Internet Research Group (QIRG) [99], which group is formulated and supported by an international researcher background and collaboration. The authors of [98] address some important capability milestones for the realization of a global-scale quantum Internet. The technical roadmap also addresses important future engineering problems brought up by the quantum Internet, such as the development of a standardized architectural framework for the quantum Internet, standardization and protocols of the quantum Internet,

layer interoperability, advanced services for the quantum Internet, interoperability of the traditional Internet and quantum Internet, connection establishment between the heterogeneous quantum nodes of the quantum Internet, definition of node roles, network coding, multiparty state transfer, entanglement distribution mechanisms and entanglement routing, application programming interface (API) for the quantum Internet, and the definition of the application level of the quantum Internet.

In [100], the authors defined a method for deterministic delivery of quantum entanglement on a quantum network. The results allow us to realize entanglement distribution across multiple remote quantum nodes in a quantum Internet setting.

In [45], the authors demonstrated the quantum teleportation of independent single-photon qubits over 1,400 kilometres. Since an experimental realization of a global-scale quantum Internet requires the application of quantum teleportation over long-distances, the proposed results represent a fundamental of any experimental quantum Internet. In [46], the authors demonstrated quantum teleportation with high fidelity values between remote single-atom quantum memories.

Some other recent results connected to the development of an experimental global-scale quantum Internet are as follows. In [101], the authors demonstrated the Bell inequality violation using electron spins separated by 1.3 kilometres. In [102], the authors demonstrated modular entanglement of atomic qubits using photons and phonons. The quantum repeaters are fundamental networking elements of any experimental quantum Internet. The quantum repeaters are used in the entanglement distribution process to generate quantum entanglement between distant senders and receivers. The quantum repeaters also realize the entanglement purification (entanglement improvement) and the entanglement swapping (entanglement extension) procedures. For an experimental realization of quantum repeaters based on atomic ensembles and linear optics, see [103].

Since quantum channels also have a fundamental role in the quantum Internet, we suggest the review paper of [29], and also the work of [104], for some specialized applications of quantum channels. For a review on some recent results of quantum computing technology, we suggest [105]. Some recent services developed for the quantum Internet can be found in [112]–[116]. The works [91]–[93], [96] are related to the utilization of entanglement for long-distance quantum communications and for a global-scale quantum Internet, and also to the various aspects of quantum networks in a quantum Internet setting.

For some fundamental works on quantum Shannon theory, see [27]–[29], [104], [106]–[109]. For some important works on the experimental implementations of quantum repeaters, entanglement purification and entanglement distribution, see [110]–[112], [117]–[119].

V. CONCLUSION

Here we provided a brief overview of the recent results of QKD. The review focused on the principles of DVQKD and CVQKD protocols, the main attributes of the recent implementations, and the integration of QKD into traditional and quantum communication networks.

REFERENCES

[1] Skorin-Kapov, N. et al., Physical-Layer Security in Evolving Optical Networks, *IEEE Commun. Mag.*, vol. 54, no. pp. 110–117. (2016).
[2] Bennett, C. H. and Brassard, G. Quantum cryptography: Public-key distribution and coin tossing. In: *Proceeding of the IEEE International Conference on Computers Systems and Signal Processing*. Washington: IEEE, pp. 175-179 (1984).
[3] Ekert, A. K. Quantum cryptography based on Bell's theorem. *Phys Rev Lett*, 67, pp. 661-663 (1991).
[4] Yuen, H. P. Security of quantum key distribution. *IEEE Access*, Vol.4, No.1, pp. 724-749 (2016).
[5] Broadbent, A., and Schaffner, C. Quantum cryptography beyond quantum key distribution. *Designs Codes and Cryptography*, Vol.78, No.1, pp. 351-382 (2016).
[6] Cao, Y. et al. Key as a Service (KaaS) over Quantum Key Distribution (QKD)-Integrated Optical Networks, *IEEE Comm. Mag.* DOI: 10.1109/MCOM.2019.1701375 (2018).
[7] Lo, H.-K. et al., Secure Quantum Key Distribution, *Nature Photon.*, vol. 8, pp. 595–604. (2014)
[8] Mao, Y. et al., Integrating Quantum Key Distribution with Classical Communications in Backbone Fiber Network, *Opt. Express*, vol. 26, no. 5, pp. 6010–6020. (2018).
[9] Karinou, F. et al., Toward the Integration of CV Quantum Key Distribution in Deployed Optical Networks, *IEEE Photon. Technol. Lett.*, vol. 30, no. 7, pp. 650–653. (2018).
[10] Frohlich, B. et al., Long-Distance Quantum Key Distribution Secure Against Coherent Attacks, *Optica*, vol. 4, no. 1, pp. 163–167. (2017).
[11] Wang, L.-J. et al., Long-Distance Copropagation of Quantum Key Distribution and Terabit Classical Optical Data Channels, *Phys. Rev. A*, vol.95, no. 1, pp. 012301. (2017).
[12] Cao, Y. et al., Key on Demand (KoD) for Software-Defined Optical Networks Secured by Quantum Key Distribution (QKD), *Opt. Express*, vol.25, no. 22, pp. 26453–26467 (2017).
[13] Cao, Y. et al., Time-Scheduled Quantum Key Distribution (QKD) over WDM Networks, *J. Lightwave Technol.*, vol. 36, no. 16, pp.3382–3395. (2018).
[14] Peev, M. et al., The SECOQC Quantum Key Distribution Network in Vienna, *New J. Phys.*, vol. 11, no. 7, pp. 075001. (2009).
[15] Aguado, A. et al., Secure NFV Orchestration over an SDN-Controlled Optical Network with Time-Shared Quantum Key Distribution Resources, *J. Lightwave Technol.*, vol. 35, no. 8, pp. 1357–1362. (2017).
[16] Xu, S. et al. Fiber-Wireless Network Virtual Resource Embedding Method Based on Load Balancing and Priority, *IEEE Access*, vol. 6, pp. 33201-33215 (2018).
[17] Lucamarini, M. et al., Overcoming the Rate–Distance Limit of Quantum Key Distribution without Quantum Repeaters, *Nature*, vol. 557, no.7705, pp. 400–403. (2018).
[18] Comandar, L. C. et al. Room temperature singlephoton detectors for high bit rate quantum key distribution. *Appl. Phys. Lett.* 104, 021101 (2014).
[19] Yin, H.-L. et al. Measurement-device-independent quantum key distribution over a 404 km optical fiber. *Phys. Rev. Lett.* 117, 190501 (2016).
[20] Aleksic, S. et al., Towards a Smooth Integration of Quantum Key Distribution in Metro Networks, in *Proc. 16th Int. Conf. on Transparent Optical Networks* (ICTON), Graz, Austria, (2014).
[21] Zhao, Y. et al., Energy Efficiency with Sliceable Multi-Flow Transponders and Elastic Regenerators in Survivable Virtual Optical Networks, *IEEE Trans. Commun.*, vol. 64, no. 6, pp. 2539.2550. (2016).
[22] Jouget P. et al. Experimental demonstration of long-distance continuous-variable quantum key distribution, *Nature Photonics* volume 7, pages 378–381 (2013).
[23] Galambos, M. and Bacsardi, L. Comparing Calculated and Measured Losses in a Satellite-Earth Quantum Channel, *Infocomm. Journal* X:3, pp. 14-19., (2018).
[24] Galambos, M. and Bacsardi, L. The Evolution of Free-Space Quantum Key Distribution, *Infocomm. Journal* X:1 pp. 22-30.,(2018).
[25] Bacsardi, L. On the Way to Quantum-Based Satellite Communication, *IEEE Comm. Mag.* 51:(08) pp. 50-55. (2013).
[26] Gyongyosi, L. and Imre, S. Geometrical Analysis of Physically Allowed Quantum Cloning Transformations for Quantum Cryptography, *Information Sciences*, Elsevier, pp. 1-23, DOI: 10.1016/j.ins.2014.07.010 (2014).
[27] Imre, S. and Gyongyosi, L. *Advanced Quantum Communications - An Engineering Approach*. New Jersey, Wiley-IEEE Press (2013).
[28] Petz, D. *Quantum Information Theory and Quantum Statistics*, Springer-Verlag, Heidelberg, Hiv: 6. (2008).

[29] Gyongyosi, L., Imre, S. and Nguyen, H. V. A Survey on Quantum Channel Capacities, *IEEE Communications Surveys and Tutorials*, DOI: 10.1109/COMST.2017.2786748 (2018).

[30] Shor, P. W. Algorithms for quantum computation: Discrete logarithms and factoring. In *Proc. 35th Symposium on Foundations of Computer Science*, 124–134, Los Alamitos, CA, IEEE Computer Society Press (1994).

[31] IBM. *A new way of thinking: The IBM quantum experience*. URL: http://www. Research.ibm. Com/quantum. (2017).

[32] Monz, T. et al. Realization of a scalable Shor algorithm. *Science* 351, 1068-1070 (2016).

[33] Vandersypen, L. M. K. et al. Experimental realization of Shor's quantum factoring algorithm using nuclear magnetic resonance. *Nature* 414, 883-887 (2001).

[34] Aaronson, S. and Chen, L. Complexity-theoretic foundations of quantum supremacy experiments. *Proceedings of the 32nd Computational Complexity Conference*, CCC '17, pages 22:1-22:67, (2017).

[35] Harrow, A. W. and Montanaro, A. Quantum Computational Supremacy, *Nature*, vol 549, pages 203-209 (2017).

[36] Preskill, J. Quantum Computing in the NISQ era and beyond, *Quantum* 2, 79 (2018).

[37] Wootters, W. and Zurek, W. H. A single quantum cannot be cloned. *Nature*, 299:802–803, doi:10.1038/299802a0. (1982).

[38] Zhao, W., Liao, Q., Huang, D. et al. Performance analysis of the satellite-to-ground continuous-variable quantum key distribution with orthogonal frequency division multiplexed modulation, *Quant. Inf. Proc.* 18: 39. DOI: 10.1007/s11128-018-2147-8 (2019).

[39] Liao, S.-K. et al. Satellite-to-ground quantum key distribution, *Nature* 549, pages 43–47, (2017).

[40] Bennett, C., Quantum cryptography using any two nonorthoganol states, *Phys. Rev. Lett.* 68, pp. 3121-3124. (1992).

[41] Kwiat, P. G. Hyper-entangled states. *J Mod Opt* 44, pp. 2173-2184 (1997).

[42] Klaus, M. et al. Dense coding in experimental quantum communication. *Phys. Rev. Lett.*, 69, pp. 4656-4659 (1992).

[43] Bennett, C. H., Brassard, G., Crepeau, C., et al. Teleporting an unknown quantum state via dual classical and Einstein-Podolsky-Rosen channels. *Phys Rev Lett*, 70, pp. 1895-1899 (1993).

[44] Bouwmeester, D., Pan, J. W., Mattle, K. et al. Experimental quantum teleportation. *Nature*, 390, pp. 575-579 (1997).

[45] Ren, J.-G. et al. Ground-to-satellite quantum teleportation, *Nature* 549, pages 70–73, (2017).

[46] Noelleke, C. et al, Efficient Teleportation Between Remote Single-Atom Quantum Memories, *Physical Review Letters* 110, 140403, (2013).

[47] Hillery, M., Buzek, V. and Berthiaume, A. Quantum secret sharing. *Phys Rev A*, 59, pp. 1829-1834 (1999).

[48] Jiang, Y. et al. Quantum secret sharing protocol and its modeling checking. *Laser and Optoelectronics Progress* 54(12), 122704, (2017).

[49] Kiktenko, E. O., Pozhar, N. O., Anufriev, M. N., Trushechkin, A. S., Yunusov, R. R., Kurochkin, Y. V., Lvovsky, A. I. and Fedorov, A. K. Quantum-secured blockchain, *Quantum Sci. Technol.* 3, 035004 (2018).

[50] Grosshans, F., Cerf, N. J., Wenger, J., Tualle-Brouri, R. and Grangier, P. Virtual entanglement and reconciliation protocols for quantum cryptography with continuous variables. *Quant. Info. and Computation* 3, 535-552 (2003).

[51] Navascues, M. and Acin, A. Security bounds for continuous variables quantum key distribution. *Phys. Rev. Lett*. 94, 020505 (2005).

[52] Pirandola, S., Mancini, S., Lloyd, S. and Braunstein, S. L. Continuous-variable Quantum Cryptography using Two-Way Quantum Communication, *Nature Physics* 4, 726 - 730 (2008).

[53] Pirandola, S., Garcia-Patron, R., Braunstein, S. L. and Lloyd, S. *Phys. Rev. Lett.* 102 050503. (2009).

[54] Pirandola, S., Serafini, A. and Lloyd, S. *Phys. Rev. A* 79 052327. (2009).

[55] Pirandola, S., Braunstein, S. L. and Lloyd, S. *Phys. Rev. Lett.* 101 200504 (2008).

[56] Weedbrook, C., Pirandola, S., Lloyd, S. and Ralph, T. *Phys. Rev. Lett.* 105 110501 (2010).

[57] Weedbrook, C., Pirandola, S., Garcia-Patron, R., Cerf, N. J., Ralph, T., Shapiro, J. and Lloyd, S. *Rev. Mod. Phys.* 84, 621 (2012).

[58] Shieh, W. and Djordjevic, I. *OFDM for Optical Communications*. Elsevier (2010).

[59] Navascues, M., Grosshans, F. and Acin, A. Optimality of Gaussian Attacks in Continuous-variable Quantum Cryptography, *Phys. Rev. Lett.* 97, 190502 (2006).

[60] Garcia-Patron, R. and Cerf, N. J. Unconditional Optimality of Gaussian Attacks against Continuous-Variable Quantum Key Distribution. *Phys. Rev. Lett.* 97, 190503 (2006).

[61] Grosshans, F. Collective attacks and unconditional security in continuous variable quantum key distribution. *Phys. Rev. Lett.* 94, 020504 (2005).

[62] Laudenbach, F., Pacher, C., Fred Fung, C-H., Poppe, A., Peev, M., Schrenk, B., Hentschel, M., Walther, P. and Hubel, H. Continuous-Variable Quantum Key Distribution with Gaussian Modulation - The Theory of Practical Implementations, *Adv. Quantum Technol. 1800011* (2018).

[63] Diamanti, E. and Leverrier, A. Distributing Secret Keys with Quantum Continuous Variables: Principle, Security, *Entropy*, 17, 6072-6092; doi:10.3390/e17096072 and Implementations (2015).

[64] Stucki, D. et al. High speed coherent one-way quantum key distribution prototype, *Optics Express*, Vol. 17, Issue 16, pp. 13326-13334 (2009).

[65] Inoue, K., Takesue, H. and Honjo, T. DPS quantum key distribution and related technologies, *SPIE Proceedings* Vol 7236, Quantum Communications Realized II; 72360I (2009).

[66] Enzer, D., Hadley, P., Gughes, R., Peterson, C. and Kwiat, P., Entangled-photon six-state quantum cryptography, *New Journal of Physics*, pp 45.1-45.8 (2002).

[67] Lo, H., Ma, X. and Chen, K. Decoy state quantum key distribution, *Phys. Rev. Lett*. 94, 230504, http://arxiv.org/pdf/quant-ph/0411004 (2005).

[68] Gyongyosi, L. and Imre, S. Adaptive multicarrier quadrature division modulation for long-distance continuous-variable quantum key distribution, *Proc. SPIE 9123, Quantum Information and Computation XII*, 912307; doi:10.1117/12.2050095, From Conference Volume 9123, Quantum Information and Computation XII, Baltimore, Maryland, USA (2014).

[69] Gyongyosi, L. and Imre, S. Secret Key Rate Proof of Multicarrier Continuous-Variable Quantum Key Distribution, *Int. J. Commun. Syst*. (Wiley), DOI: 10.1002/dac.3865, (2018).

[70] Gyongyosi, L. and Imre, S. Multiple Access Multicarrier Continuous-Variable Quantum Key Distribution, *Chaos, Solitons and Fractals*, Elsevier, DOI: 10.1016/j. Chaos.2018.07.006, ISSN: 0960-0779, (2018).

[71] Gyongyosi, L. and Imre, S. Gaussian Quadrature Inference for Multicarrier Continuous-Variable Quantum Key Distribution, *Quantum Studies: Mathematics and Foundations*, Springer Nature, DOI: 10.1007/s40509-019-00183-9, (2019).

[72] Gyongyosi, L. and Imre, S. Diversity Space of Multicarrier Continuous-Variable Quantum Key Distribution, *Int. J. Commun. Syst.* (Wiley), ISSN: 1099-1131, (2019).

[73] Zhang, H., Mao, Y., Huang, D., Li, J., Zhang, L. and Guo, Y. Security analysis of orthogonal-frequency-division-multiplexing-based continuous-variable quantum key distribution with imperfect modulation, *Phys. Rev. A* 97, 052328 (2018).

[74] Jouguet, P. Kunz-Jacques, S. and Leverrier, A. Long-distance continuous-variable quantum key distribution with a Gaussian modulation, *Phys. Rev. A* 84, 062317. (2011).

[75] Renner, R. and Cirac, J. I. de Finetti Representation Theorem for Infinite-Dimensional Quantum Systems and Applications to Quantum Cryptography, *Physcal Review Letters* 102, 110504 (2009).

[76] Leverrier, A., García-Patron, R., Renner, R. and Cerf, N. J. Security of Continuous-Variable Quantum Key Distribution Against General Attacks, *Physcal Review Letters* 110, 030502. (2013).

[77] Gyongyosi, L. and Imre, S. Low-Dimensional Reconciliation for Continuous-Variable Quantum Key Distribution, *Appl. Sci.,* doi: 10.3390/app8010087, ISSN 2076-3417, (2018).

[78] Vernam, G. S. Cipher printing telegraph systems for secret wire and radio telegraphic communications, *Transactions of the American Institute of Electrical Engineers*, Vol. XLV, No.2, pp. 295-301 (1926).

[79] Shannon, C. E. Communication theory of secrecy systems. *Bell Syst Technol J*, 28, pp. 656-715 (1949).

[80] Van Meter, R. *Quantum Networking*. ISBN 1118648927, 9781118648926, John Wiley and Sons Ltd (2014).

[81] Lloyd, S., Shapiro, J. H., Wong, F. N. C., Kumar, P., Shahriar, S. M. and Yuen, H. P. Infrastructure for the quantum Internet. *ACM SIGCOMM Computer Communication Review*, 34, 9–20 (2004).

[82] Kimble, H. J. The quantum Internet. *Nature*, 453:1023–1030 (2008).

[83] Caleffi, M., Cacciapuoti, A. S. and Bianchi, G. Quantum Internet: from Communication to Distributed Computing, *aXiv:1805.04360* (2018).

[84] Castelvecchi, D. The quantum internet has arrived, *Nature*, News and Comment, https://www. Nature. Com/articles/d41586-018-01835-3, (2018).

[85] Cacciapuoti, A. S., Caleffi, M., Tafuri, F., Cataliotti, F. S., Gherardini, S. and Bianchi, G. Quantum Internet: Networking Challenges in Distributed Quantum Computing, *arXiv:1810.08421* (2018).

[86] Caleffi, M. End-to-End Entanglement Rate: Toward a Quantum Route Metric, 2017 *IEEE Globecom*, DOI: 10.1109/GLOCOMW.2017.8269080, (2018).
[87] Caleffi, M. Optimal Routing for Quantum Networks, *IEEE Access*, Vol 5, DOI: 10.1109/ACCESS.2017.2763325 (2017).
[88] Liao, S.-K., Wen-Qi, C., Handsteiner, J. et al. Satellite-Relayed Intercontinental Quantum Network. *Phys. Rev. Lett*. 120, 030501, (2018).
[89] Muralidharan, S., Kim, J., Lutkenhaus, N., Lukin, M. D. and Jiang. L. Ultrafast and Fault-Tolerant Quantum Communication across Long Distances, *Phys. Rev. Lett*. 112, 250501 (2014).
[90] Rozpedek, F., Schiet, T., Thinh, L., Elkouss, D., Doherty, A., and S. Wehner, Optimizing practical entanglement distillation, *Phys. Rev. A* 97, 062333 (2018).
[91] Van Meter, R., Ladd, T. D., Munro, W. J. and Nemoto, K. System Design for a Long-Line Quantum Repeater, *IEEE/ACM Transactions on Networking* 17(3), 1002-1013, (2009).
[92] Van Meter, R., Satoh, T., Ladd, T. D., Munro, W. J. and Nemoto, K. Path selection for quantum repeater networks, *Networking Science*, Volume 3, Issue 1–4, pp 82–95, (2013).
[93] Pirandola, S., Laurenza, R., Ottaviani, C. and Banchi, L. Fundamental limits of repeaterless quantum communications, *Nature Communications*, 15043, doi:10.1038/ncomms15043 (2017).
[94] Muralidharan, S., Kim, J., Lutkenhaus, N., Lukin, M. D. and Jiang. L. Ultrafast and Fault-Tolerant Quantum Communication across Long Distances, *Phys. Rev. Lett*. 112, 250501 (2014).
[95] Bacsardi, L. On the Way to Quantum-Based Satellite Communication, *IEEE Comm. Mag.* 51:(08) pp. 50-55. (2013).
[96] Van Meter, R. and Devitt, S. J. Local and Distributed Quantum Computation, *IEEE Computer* 49(9), 31-42 (2016).
[97] Pirandola, S. Capacities of repeater-assisted quantum communications, *arXiv:1601.00966* (2016).
[98] Wehner, S., Elkouss, D., and R. Hanson. Quantum internet: A vision for the road ahead, *Science* 362, 6412, (2018).
[99] Quantum Internet Research Group (QIRG), web: https://datatracker.ietf.org/rg/qirg/about/ (2018).
[100] Humphreys, P. et al., Deterministic delivery of remote entanglement on a quantum network, *Nature* 558, (2018).
[101] Hensen, B. et al., Loophole-free Bell inequality violation using electron spins separated by 1.3 kilometres, *Nature* 526, (2015).
[102] Hucul, D. et al., Modular entanglement of atomic qubits using photons and phonons, *Nature Physics* 11(1), (2015).
[103] Sangouard, N. et al., Quantum repeaters based on atomic ensembles and linear optics, *Reviews of Modern Physics* 83, 33, (2011).
[104] Pirandola, S., Braunstein, S. L., Laurenza, R., Ottaviani, C., Cope, T. P. W., Spedalieri, G. and Banchi, L. Theory of channel simulation and bounds for private communication, *Quantum Sci. Technol*. 3, 035009 (2018).
[105] Gyongyosi, L. and Imre, S. A Survey on Quantum Computing Technology, *Computer Science Review*, Elsevier, DOI: 10.1016/j. Cosrev.2018.11.002, ISSN: 1574-0137, (2018).
[106] Lloyd, S. Capacity of the noisy quantum channel. *Physical Rev. A*, 55:1613–1622 (1997).
[107] Gisin, N. and Thew, R. Quantum Communication. *Nature Photon.* 1, 165-171 (2007).
[108] Leung, D., Oppenheim, J. and Winter, A. *IEEE Trans. Inf. Theory* 56, 3478-90. (2010).
[109] Kobayashi, H., Le Gall, F., Nishimura, H. and Rotteler, M. Perfect quantum network communication protocol based on classical network coding, *Proceedings of 2010 IEEE International Symposium on Information Theory* (ISIT) pp 2686-90. (2010).
[110] Goebel, A. M., Wagenknecht, G., Zhang, Q., Chen, Y., Chen, K., Schmiedmayer, J. and Pan, J. W. Multistage Entanglement Swapping. *Phys. Rev. Lett.* 101, 080403 (2008).
[111] Xiao, Y. F., Gong, Q. Optical microcavity: from fundamental physics to functional photonics devices. *Science Bulletin*, 61, 185-186 (2016).
[112] Gyongyosi, L. and Imre, S. Decentralized Base-Graph Routing for the Quantum Internet, *Physical Review A*, American Physical Society, DOI: 10.1103/PhysRevA.98.022310 (2018).
[113] Gyongyosi, L. and Imre, S. Multilayer Optimization for the Quantum Internet, *Scientific Reports*, Nature, DOI:10.1038/s41598-018-30957-x, (2018).
[114] Gyongyosi, L. and Imre, S. Entanglement Availability Differentiation Service for the Quantum Internet, *Scientific Reports*, Nature, (DOI:10.1038/s41598-018-28801-3), https://www.nature.com/articles/s41598-018-28801-3 (2018).
[115] Gyongyosi, L. and Imre, S. Entanglement-Gradient Routing for Quantum Networks, *Scientific Reports*, Nature, (DOI:10.1038/s41598-017-14394-w), https://www.nature.com/articles/s41598-017-14394-w (2017).
[116] Gyongyosi, L. and Imre, S. Opportunistic Entanglement Distribution for the Quantum Internet, *Scientific Reports*, Nature, DOI:10.1038/s41598-019-38495-w, (2019).
[117] Chou, C., Laurat, J., Deng, H., Choi, K. S., de Riedmatten, H., Felinto, D. and Kimble, H. J. Functional quantum nodes for entanglement distribution over scalable quantum networks. *Science*, 316(5829):1316–1320 (2007).
[118] Zhang, W. et al. Quantum Secure Direct Communication with Quantum Memory. *Phys. Rev. Lett.* 118, 220501 (2017).
[119] Rozpedek, F., Schiet, T., Thinh, L., Elkouss, D., Doherty, A., and S. Wehner, Optimizing practical entanglement distillation, *Phys. Rev. A* 97, 062333 (2018).

Laszlo Gyongyosi received degrees from the Budapest University of Technology and Economics (BME). He receives the D.Sc. degree from the Hungarian Academy of Sciences (MTA) in 2019. His research interests include quantum computation and communications, quantum information, and quantum Shannon theory. He is a research scientist at the Department of Networked Systems and Services at the BME, in contribution with the University of Southampton (Soton), U.K., and Hungarian Academy of Sciences.

Sandor Imre (M'93–SM'12) received the Dr.Univ. degree in probability theory and statistics in 1996, the Ph.D. degree in 1999, and the D.Sc. degree from the Hungarian Academy of Sciences in 2007. He is a Professor and the Head of the Department of Networked Systems and Services at the Budapest University of Technology (BME). He is chairman of Telecommunication Scientific Committee of Hungarian Academy of Sciences. He was invited to join the Mobile Innovation Centre as R&D Director in 2005. His research interests include mobile and wireless systems, quantum computing, and communications. He has contributed to different wireless access technologies, mobility protocols and their game theoretical approaches, reconfigurable systems, and quantum computing based algorithms and protocols.

Laszlo Bacsardi received his M.Sc. degree in 2006 in Computer Engineering from the Budapest University of Technology and Economics (BME). He wrote his PhD thesis on the possible connection between space communications and quantum communications at the BME Department of Telecommunications in 2012. From 2009, he works at the University of Sopron, Hungary (formerly known as University of West Hungary). He holds an associate professor position at the Institute of Informatics and Economics, University of Sopron. He is Research Fellow at the Department of Networked Systems and Services, BME. His current research interests are quantum computing, quantum communications and ICT solutions developed for Industry 4.0. He is the Vice President of the Hungarian Astronautical Society (MANT), which is the oldest Hungarian non-profit space association founded in 1956. Furthermore, he is member of IEEE, AIAA and the HTE as well as alumni member of the UN established Space Generation Advisory Council (SGAC). In 2017, he won the IAF Young Space Leadership Award.

Visible Light Communication Survey

Eszter Udvary, *Member, IEEE*

***Abstract*— Communication applying visible light technology is a novel approach. Visible Light Communication (VLC) development is motivated by the increasing demand for wireless communication technologies. It has the potential to provide high-speed data communication with good security and improved energy efficiency. The rapid evolution of VLC was sustained by the LEDs performances. The Light-Emitting-Diode (LED) luminaires are capable of switching to the different light intensity at a fast rate. This function can be used for data transmission. This article focuses on the physical layer of the VLC links. It reviews the technology, the topology of the proposed connection, and the benefits of this approach. The main research trends are identified emphasizing state of the art in this area. It shows how VLC technology evolved and what are the performances achieved at this time. Various structures of the transmitter and receiver are studied, and different modulation schemes are investigated. Finally, numerous applications of VLC technology are presented.**

***Index Terms*—Visible light communication, Optical-wireless communication, Free-space optical communication, Optical communication equipment, Modulation techniques, Machine-to-machine communications, Light emitting diodes, Lighting, Diode lasers**

I. INTRODUCTION

NOWADAYS, a growing increase in the traffic carried by the telecommunication networks, including the wireless networks, can be observed [1]. The novel bandwidth-hungry applications increase the demand for broadband internet services, and further innovation, research, and development in the new emerging communication technologies are needed. The required capacity of wireless data transmission is expected to increase exponentially in the next years.

Radio frequency (RF) type communications are applied for wireless links, because of its maturity level and full acceptance. However, the radio frequency based wireless communications have some limitations. The reliability and the performances of the link are determined by the limited available spectrum and the increasing number of nodes. The main disadvantage is the limited bandwidth. There are also some scenarios where the RF caused interferences are critical, such as in aircraft, airports, or hospitals. So, novel wireless communication technologies are required.

Meanwhile, the development of the LEDs had massive growth. The revolution in the field of solid-state lighting leads to the replacement of fluorescent lamps by Light Emitting Diodes. Nowadays, LEDs are energy efficient, highly reliable, and have a lifetime that exceeds by far the traditional light sources. So, LEDs are used in more and more lighting applications, because of the numerous advantages, and it is considered that LEDs will completely replace the traditional lighting sources [1] - [6]. On the other hand, LEDs can be used not only for lighting but also for communication, because the light intensity can be varied, and the switching speed is high enough.

Visible light communication is a new wireless communication technology which uses the white light not just for illumination purposes but also as a carrier for digital transmission. VLC uses the visible light (frequency range=430-790THz, wavelength range=380-750nm) as a communication medium, which offers enormous bandwidths free of charge, this frequency range is safe to the human body, it does not disturb any sensitive electrical equipment, the allowed power is high, and it is not limited by any law, because of the applied non-licensed frequency range. As a visible light source can be used both for illumination and communication; therefore, it saves the extra power that is required in RF communication. The applied LEDs are energy efficient, small size, and cost-effective.

The basic concept is simple; the information modulates the intensity of the VLC transmitter. At the receiver side, a photosensitive element extracts the data from the detection of the fluctuation of the light intensity. The main advantage of VLC system is the application of the multifunctional device, which is used for lighting and data transmission same time. The communication link uses the existing LEDs lighting systems. With this approach, the implementation cost of the transmission link is significantly reduced.

Additionally, VLC link offers a considerable bandwidth available free of charge, enabling high data rate communications without any RF interference. VLC technology can provide low-cost, high-speed, optical-wireless data communication. VLC is a new technology, but the development is fast.

This paper aims at providing a survey to the physical layer of VLC technology. It presents the architecture of a VLC system, overviews the advantages and the disadvantages of the technology. This survey focuses on the applied modulation methods in the VLC systems. It identifies and discusses several top applications of VLC, pointing out the benefits of VLC usage. This article does not aim with the detailed VLC channel modeling and the standardization of VLC technology. The organization of this paper is as follows. Section II describes the architecture of VLC systems. Section III overviews the description of the potential applications of VLC. Section IV presents the state of the art of technology. Finally, section VI concludes the paper.

Eszter Udvary is an Associate Professor at Budapest University of Technology and Economics, Budapest, Hungary (e-mail: udvary@hvt.bme.hu).

II. VISIBLE LIGHT COMMUNICATION LINK

The VLC system consists of a VLC transmitter that modulates the white light produced by LEDs; a VLC receiver based on a photodiode that extracts the modulated signal from the light power, and the VLC wireless optical channel to connect the physically separated VLC transmitter and receiver. The simplified block diagram of a VLC system is presented in Fig 1.

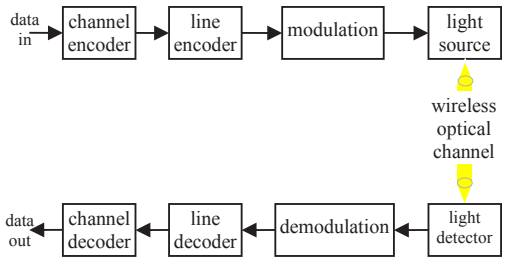


Fig.1. VLC link structure

A. The transmitter side

VLC transmitter transforms data into messages that can be sent over the free space optical medium by using visible light. The primary purposes that it is a multifunctional device; it emits light and transmits data at the same time. On the transmitter side, white light is generated by the LED, and the light is modulated by the information. The data transmission must not affect the primary illumination function of the device. From this point of view, the VLC transmitter must be met with the lighting requirements. So, the same optical power is used, or it is allowed for dimming. The dimming level that is selected for the modulation should be such that it is supported by the illuminating LEDs. On the other hand, the VLC transmitter must not induce any noticeable flickering. The modulation should be done in a way to avoid flickering.

Two types of white-light sources are used in solid-state lighting. Red-green-blue (RGB) emitter provides the white light applying three colors. The blue-LED on yellow-light emitting phosphorus layer provides white light by mixing blue and yellow lights. The VLC data transmitter can use both types, but RGB solution gets more modulation bandwidth. Contrary, the blue LED based device is more energy efficient and lower complex. Based on it, a blue LED with phosphorus is more popular in illumination systems. The RGB approach can be improved by applying a fourth color. RGBY model is supplemented by the yellow color, and therefore, there is not necessary to create complicated combinations of the fundamental wavelengths. As a result, the bitrate of the system is increased, and communication is possible on four independent channels. This fourth channel also improves area coverage [7].

The parameters of the VLC transmitter are mainly limited by the characteristics of the LEDs. The data rate depends on the switching abilities of the LEDs. The service area depends on the transmission power and the illumination angle. Currently, the industry produces LEDs that can offer switching frequencies of a few tens of megahertz. The modulation bandwidth is about 2.5 MHz for the white component generated by a blue LED with yellow phosphorus. The switching speed of the blue LEDs is better, and higher data rates are enabled [8]. So, the modulation bandwidth can be increased by the filtering out of the yellow element in the receiver when only the blue part is detected. If this filtering eliminates the slow response of the yellow phosphorus, 14 MHz bandwidth can be achieved. Several other approaches are proposed to increase the bandwidth. Fully integrated LED driver design can provide high speed, low size, and economical power consumption solution. The 3 dB bandwidth of a VLC transmitter can be extended to 80 MHz applying an integrated driver circuit with high pass transfer function [9].

To achieve higher VLC data rate, LASER (Light Amplification by Stimulated Emission of Radiation) diode transmitter has been proposed and demonstrated [10]. In this approach, the main challenge is the contemporary lighting and communication features. Nowadays, it is not applied in the lighting system.

Different types and forms of LED are applied in various environments. High power LEDs or LED arrays are used in illustrative in-door illumination purposes. Low-power devices are utilized in smart-phones and other mobile devices.

B. Modulation techniques

At the transmitter side, a dimming or biasing circuit with control function is necessary. Application of microcontrollers is a cost-effective solution for the encoder. The microcontroller can be replaced by a Field Programmable Gate Array (FPGA) in more complex applications. FPGA provides enhanced performances with the help of digital signal processing techniques. The encoder in the transmitter converts the data into a modulated message and manages the switching of the LEDs according to the binary data and the imposed data rate. So, the binary data are converted into an intensity modulated light beam. The visible light communication systems use intensity modulation and direct detection (IM/DD) approach. For IM/DD systems, the optical intensity must be real-valued and non-negative. As a result of the constraints of IM/DD, modulation schemes that are advantageous in radio frequency communications that may not be offered the same advantage in VLC. Typically, the light produced by the LEDs is current modulated with a baseband modulation signal. The baseband modulation schemes, like various type of On-Off Keying (OOK) modulation, Pulse Amplitude Modulation (PAM), and Pulse Position Modulation (PPM), are often applied [11]. Multicarrier modulation techniques as Orthogonal Frequency Differential Modulation (OFDM) can be used to increase the system capacity [12]. A unique modulation method, called Color Shift Keying (CSK) is specially designed for visible light communication to overcome the low data rate [13].

The baseband modulation schemes can be classified into pulse amplitude, pulse position, and pulse interval modulation depending upon the method information is encoded into the optical carrier. In On-Off Keying, the LEDs are turned off and

on according to the bits in the stream; it is the same approach as the applied modulation in fiber optic systems. Typically, “1” bit is represented by the on state and "0" bit is represented by the off state. The LED is not turned completely off in the off state, because of the better modulation performance. The power requirement steadily decreases as the duty cycle decreases, but the bandwidth requirement increases. The implementation of OOK is easy and straightforward. The application of on-off keying modulation is limited by the slow time response of the yellow phosphor case of the blue emitter and yellow phosphor structure. It defines the modulation bandwidth. Typically, data rate up to 10 Mbps can be realized using NRZ (Non-Return-to-Zero) OOK with white LED [14]. The data transmission rate can be improved with analog equalization, integrated driving circuits, and blue filtering techniques on the receiving side [9].

The low data rate OOK motivated researchers to develop new modulation techniques to achieve higher data modulation rates. In the case of Pulse Width Modulation (PWM), the width of the pulses varies according to dimming levels. The different dimming levels can be varied between 0% and 100% by applying high PWM frequency.

As the name suggests, the information in the Pulse Positioning Modulation scheme is encoded in the position of a pulse within a symbol. An L-PPM symbol consists of L time slots of equal duration. Within the symbol, all slots except the information bearing slot are empty. The position of this pulse carries information about the input bit sequence. The location of the pulse corresponds to the decimal value of the M-bit input data. So, a single pulse is presented in each symbol period; this scheme suffers from the problem of the low data rate. For smaller values of ‘L’, it is not efficient in terms of power and bandwidth usage. Multi-pulse PPM (MPPM) is a variant of PPM modulation schemes. It is more spectrally efficient because multiple pulses are transmitted in each symbol-time. A modified version of the PPM is the Expurgated PPM (EPPM) which was introduced to improve the performance of peak-power limited M-ary communication systems. The spectral efficiency of the MPPM and EPPM is less than 1, Multilevel EPPM (MEPPM) is proposed for the better spectral efficiency.

Color Shift Keying is a unique modulation method in VLC systems to enhance the data rate. This modulation scheme is designed to operate with RGB LEDs to provide higher order, spectrally efficient modulation. Three separate LEDs (Green, Blue, and Red) are utilized to produce the white light. Modulation in CSK is realized using the intensity of the three colors in an RGB LED source. Data are sent on the instantaneous color of the RGB triplet. CSK depends on the color space chromaticity diagram. The constant emitted light guarantees an absence of flicker at all frequencies. The constant luminous flux of the source leads to near constant current drive, which in turn implies a reduced inrush current when modulating data, strong signal isolation from the power line and a reduction in inductance caused by large switching currents. The bit rate is decided by the symbol rate and the number of color points on the constellation. That means the frequency response of the LEDs does not limit CSK bit rate. The main disadvantage of this approach, which Phosphor-based visible LEDs are more often used, and they are not suitable for CSK.

The VLC link has two main challenges: the limited bandwidth of the LEDs and the multipath propagation. The typical modulation bandwidth of LEDs is around couple tens of MHz. Complex modulation schemes such as phase shift keying (PSK), quadrature amplitude modulation (QAM) or OFDM modulations can be used to realize a higher data rate. The most popular and applicable choice in VLC systems is OFDM since it offers improved spectral efficiency than PSK, QAM and it has strong robustness against the intersymbol interference (ISI) airing from multipath propagation or limited system bandwidth.

High data rates exceeding 100 Mb/s are also attainable with multiple-subcarrier modulation techniques such as OFDM. With arrays of separately driven light sources and OFDM, data throughput of up to 1Gbit/s was demonstrated, applying methods similar to radio frequency multiple-input and multiple-output (MIMO) approach. Multicarrier modulation schemes can be more efficient than the baseband modulation schemes. Table I. overviews the properties of the main baseband modulation schemes.

TABLE I
PROPERTIES OF BASEBAND MODULATION SCHEMES

	Data rate	SNR	BER
OOK	medium	low	high
PWM	very low	high	low
PPM	low	high	medium
MPPM	high	medium	high
CSK	medium	medium	medium
MIMO OFDM	very high	high	low

The traditional OFDM signal widely applied to RF system is complex and bipolar. Due to IM/DD, the signaling for the VLC network must be a real and unipolar. Therefore, the traditional OFDM signal is modified to make them real-valued and unipolar. There are several variations of the unipolar OFDM that is proposed for VLC systems such as DC-biased optical OFDM (DCO-OFDM), asymmetrically clipped optical OFDM (ACO-OFDM), unipolar OFDM (U-OFDM), pulse-amplitude modulated discrete multitone modulation (PAM-DMT) and flip-OFDM. DCO-OFDM adds a DC-bias to the bipolar OFDM signal. The required DC-bias to satisfy non-negativity is equal to the maximum negative amplitude of the OFDM signal. Negative signal clipping at the zero levels is applied to realize ACO-OFDM, which improves the power efficiency of the unipolar OFDM modulation format. Since only odd subcarrier is modulated, the ACO-OFDM has only the half the spectral efficiency of DCO-OFDM. However, there is no information loss when the signal is clipped, because of the anti-symmetry of the modulated signal. Pulse amplitude modulated discrete multitoned is similar to ACO-OFDM [15], but the subcarriers are modulated by PAM. In a PAM-DMT system, there is no DC bias. All of the subcarriers are modulated, but the modulation uses only the imaginary

part of the subcarrier [16]. Thus the spectral efficiency is the same as ACO-OFDM. Although it has limited spectral efficiency, it is more power efficient than DCO-OFDM, because it also has an anti-symmetry (Hermitian symmetry). Unipolar OFDM (U-OFDM) is almost the same concept named Flip-OFDM. In this case, the negative and the positive part of the real bipolar OFDM signal are extracted. Hence, Hermitian-symmetry is preserved. The polarity of the harmful components of the symbol is inverted before the transmission of both positive and negative elements in a consecutive OFDM symbol. Table II summarizes the properties of the OFDM schemes.

TABLE II
PROPERTIES OF OFDM SCHEMES

	power efficiency	spectral efficiency	Hermitian symmetry
DCO OFDM	good	good	anti-symmetry
ACO-OFDM	improved	half	anti-symmetry
PAM-DMT flip-OFDM	improved	half	anti-symmetry

C. The receiver side

The VLC receiver extracts the data from the modulated light beam. The optical lens at the receiver collects and concentrates the incoming light to a photo-detecting element. Imaging and non-imaging receivers may be used. The detection bandwidth is typically higher than the limit of the transmitter and channel dispersion. In mobile devices, like smartphones, tablets, low-cost photodiode, or optical sensor is applied. At the output of the IM/DD link, the light power is converted to current by the photodiode. The produced current signal is proportional to the intensity of the incident wave and depends on the photodiodes spectral sensitivity.

So, the photodiode transforms the light into an electrical signal that will be demodulated and decoded by the embedded decoder module. Depending on the required performances and the cost constraints, the decoder can be a microcontroller or an FPGA. The VLC receivers are based on typically a reverse biased photodiode operating in a photoconductive mode which has high bandwidth and offers the possibility of high-speed communications. At the electrical output of the receiver, significant interference can be observed, because of the other artificial or natural light sources. An optical filter can enhance the performances of the VLC receiver. The optical filter decreases the unwanted spectrum components. Moreover, in high-speed applications using white LEDs, the optical filter allows only the passage of narrowband radiation, corresponding to the blue color.

The effect of the interferences can also be reduced by narrowing the receiver field of view (FOV), but it influences the service area. If wider FOV is applied, a more extensive service area can be supported. But more noise is captured, and the Signal to Noise Ratio (SNR) is degraded. Indoor short-range applications require increased mobility, and narrow FOV approach is not useful in these scenarios. In case of outdoor long range applications, the narrow FOV is a practical solution, because the long-range link induces small angles, anyhow. A narrow FOV improves the robustness of the VLC system again, the noise due to daylight or from other VLC transmitters [17]. The FOV of the whole receiver is determined by the FOV of the optical focusing system, which concentrates the light on the photodetector by using a lens.

Similarly, the performances of the system can be enhanced by increasing the area of the photodetector. However, if the area of the photodetector is large, its capacitance is also increased. The capacitance and the load resistor limit the available bandwidth. The applied photodetectors area represents a trade-off between SNR and bandwidth. The generated photocurrent is low, and a transimpedance circuit is used to transform the small current into voltage. The transimpedance solution offers a trade-off between the gain-bandwidth product and noise.

The output voltage of the transimpedance circuit is amplified and filtered to remove high and low-frequency noises, and the DC component. Finally, the data processing unit decodes the information from the reconstructed signal obtaining the binary message.

D. The VLC channel

The transmitter and the receiver are interconnected through the free space optical communication channel. Practically, the visible light is an electromagnetic wave; as any electromagnetic radiation, the intensity of the visible light decreases with the square root of the distance as it passes through the channel. This paper does not aim with the detailed VLC channel modeling; just a short overview is included.

An optical wave propagating in an unguided medium has to go through many constituents, which are different types in different environments. The molecules and aerosols are the main absorbers in the atmosphere. On the other hand, water molecules, chlorophyll, colored dissolved organic and suspended particulate matters; dissolved salts affect the propagation in the underwater medium. These constituents cause the optical wave to get scattered and absorbed, which in turn results in the degradation and attenuation of the received optical signal.

The VLC transmission channel is affected by numerous sources of optical noise. In the daytime, the most critical noise source is the sun. Other sources of noise are represented by other VLC transmitters or any source of light with or without data transmission capabilities. In outdoor environments, the weather causes more problems for VLC applications. The rain, snow, or dense fog includes water particles. It causes scattering of the light containing the data and affects the performance of the VLC link. The noise sources and the low signals significantly degrade the SNR in VLC, especially at a long distance link. There are several possibilities for enhancing the SNR at the receiver, like optical filtering, the adequate design of the optical system, or adaptive electrical gain and filtering.

The turbulence is the random refractive index in time and space due to the circulation of air or water, which cause fluctuation in the received signal. The turbulence is developed mainly by the variations in the temperature, pressure, and humidity. The applied formulations for turbulence power

spectra are various in different environments such as atmosphere, space, and underwater. The average intensity and the scintillation index are often calculated based on the Rytov method and the extended Huygens-Fresnel principle. Effects of different optical beam profiles, the transmitter and the receiver aperture averaging to reduce the turbulence degradation can also be taken into account [18].

In VLC link the multipath propagation is experienced mainly at indoor short transmitter-receiver distances [19] and unique scenarios as underwater and vehicular applications. Besides the LOS component, there are a large number of reflections among ceiling, walls, and floor as well as any other objects within the environment. The rays of light hit the other walls and are reflected towards the receiver. The receiver can only detect the rays entering its field of view. The detailed channel model of visible light communication system takes into account the position, size, and shape of the obstacle in the illumination area of the light source [20]. Same results can be observed within the marine environment. The reflection characteristics of the sea surface and sea bottom, as well as the presence of human and human-made objects, determine the communication [21]. In the vehicular applications, the primary reflectors are the other cars which are located in the next lanes [22].

III. Purposes Of The Visible Light Communication

A. VLC advantages

VLC can be considered to be the next generation of wireless communications, because of the unique characteristics and advantages. VLC can solve some of the problems of RF communication. The benefit of VLC comes from the benefits of the visible light. High bandwidth can be used for free of charge, which allows high data rates, unlicensed spectrum, and safety for the human body and high-precision electronic equipment. From the point of security, VLC is also better than radio frequency communication. Besides these advantages, VLC is a low-cost technology, and the implementation is straightforward as it uses the same infrastructure than a lighting system.

1) High bandwidth

The RF communications have up to 300 GHz available bandwidth, which is used for different types of applications, and consequently, the networks are often saturated. Both radio- and television broadcasting, GSM (Global System for Mobile Communications, formerly Groupe Spécial Mobile), satellite communications, military applications, etc. use the same RF frequency range. The extension of the bandwidth is costly, and sometimes it is not possible, at all. On the other hand, the complexity of the higher frequency equipment increases, and the devices are expensive. Most of the frequency ranges are licensed; the available bandwidth in unlicensed frequency ranges is limited.

VLC uses the visible light spectrum, which is between 380 and 780 THz; it adds 400 THz available bandwidth for the wireless communications. The VLC can use worldwide an unregulated and almost unlimited bandwidth offering multi-Gb/s data rates.

2) Unrestricted technology

In social environments, like hospitals or aircraft the radio frequency communication is restricted. The reason that RF communication may cause malfunctions of the high precision electronic equipment. Light communication does not affect any RF signal or equipment, because of the elimination of electromagnetic interference. So, VLC is safe for these places.

3) Security

There is no wireless communication method without risking of eavesdropping. Only quantum communication wireless channel can discover the presence of an eavesdropper, based on the photon statistic [23]. But, compared with radio frequency wireless communication approaches, the visible optical-wireless link provides higher security against eavesdropping, because the light cannot penetrate through walls. The main security limitation of indoor VLC channel is the point when the wall is open. The common weakness is the window. When the eavesdropper is in a LOS position, it and can receive the signal via the window [24]. The VLC transmitters are located near the window are more endangered. But, far VLC transmitters can also be affected via reflections from the walls and indoor equipment. On the other hand, theoretical eavesdropping of VLC based communications is a possibility through keyholes and door gaps. So, careful and adequate planning of the system is required in sensitive areas. However, VLC technique provides higher security than other radio frequency wireless links. Based on it, the VLC link is suitable in military applications or areas of high security.

4) Low implementation cost

There are three main reasons for the lower price of VLC compared with other wireless technologies. First, RF systems use a regulated band. Contrary, the visible light is in an unlicensed region of the electromagnetic spectrum. The implementation cost is significantly reduced because there is no cost for a license. Secondly, VLC will rely on existing infrastructures that are already accepted and widespread across the world. So, VLC is simple, without requiring complex modifications on the existing infrastructure. It also decreases the implementation cost. The third aspect is its reduced complexity. VLC primarily uses LED transmitters and photodiode receivers; these components are inexpensive.

5) Green wireless communication technology

The natural resource consumption and climate deteriorations are increasing as the Earth's population are growing, and human society is developing. Greenhouse gas emissions have reached alarming levels; it produces significate climate changes that affect the whole ecosystem [25]. Natural resource consumption and pollution can be significantly reduced by decreasing energy consumption. A significant percentage of energy consumption comes from artificial lighting, commonly provided by electric lights. Worldwide, approximately 20% of electricity is used for illumination, while electricity represents about 15% of the total energy produced [25].

VLC is a green wireless communication technology because it does not use additional power for communication. The same infrastructure and light are used for illuminating are also used

for data transmission. Another essential advantage of VLC is the utilization of LEDs which provide substantial energy savings, reducing the CO_2 emissions.

6) Safe for human health

The application of visible light for data transmission is entirely safe for human health. The RF adverse effect on human health is not adequately demonstrated. But, RF electromagnetic waves are currently classified as a possible cause of cancer in humans by the World Health Organization. The infra-red light (IR) is also used for wireless communications, but it has a heating effect on the incident surface. So, high power IR light can cause irreversible thermal damage of the cornea, making it harmful for the human eye [26].

B. VLC Challenges

As a conclusion of the previous sub-session, VLC is a technology that has plenty of essential advantages. Naturally, VLC has also some drawbacks. Most of the disadvantages are due to the early stage of the VLC technology. It could be overtaken as the technology is fully developed. The other ones are due to the usage of the light and its characteristics. Completely mitigate them is difficult, but their effects could be reduced, or the communication could be adapted to the situations.

1) The Line of Sight condition

VLC mostly requires line-of-sight (LoS) transmission; otherwise, the received light signal may be very weak. In a general aspect, LoS maximizes power efficiency and minimizes multipath distortion. Additionally, the interferences from other receivers are limited, and communication security is enhanced. However, Non-LoS communications are more reliable, flexible, and robust. The necessary LoS condition hurts mobility and, in some areas, it represents VLC's most significant disadvantage because an object interposed between the transmitter and receiver can block the communication unless an alternate route is available [27]. In multi-hop communications and retransmissions, the data can reach users that are located outside the transmitter's LoS but are in the serving area of another transceiver [28]. VLC system can be combined with RF links [29] when VLC cannot address a node; RF serves it.

2) Limited transmission range

VLC cannot reach the same transmission range than RF communications [30]. The VLC transmission range can be increased by optimizing the transmitter and receiver parameters, but it is still significantly shorter than the RF communication range. On the transmitter's side, the communication range can be increased by increasing the transmission power or by using a more directive light beam. So, the visible LASER diode can reach a larger distance than LED, but with this device, we lost the main advantage of the system, the LASER source cannot be used as a multifunctional transmitter. On the receiver's hand, the range can be improved by using different techniques for SNR enhancement, such as narrow FoV receiver, an optical lens or different filtering techniques. Additionally, multi-hop networking can also increase the communication range of the VLC network.

3) Susceptibility to interferences

VLC is affected by the light from other incandescent or fluorescent light sources [31]. Typically, these light sources produce low-frequency noise which can be removed with a high pass filter. In outdoor applications, the sunlight causes strong perturbation. The sun produces unmodulated light, which creates a strong DC electrical component at the output of the receiver. It can be removed with capacitive DC filter. However, high-intensity optical noise can saturate the receiver and block the communication.

The degrading effect of other light sources can be reduced by optimizing the receiver. Optical filters, reduced receiver FOV, and electrical filter can eliminate the unwanted frequency components. Even if the mentioned technique mitigates the effect of the interferences, high levels of noise still affect the communication performances.

IV. APPLICATIONS

Essential features of VLC include high bandwidth, no health hazard, low power consumption, and non-licensed channels made it attractive for practical use. By taking benefit of the mentioned advantages, VLC technology can support numerous applications. VLC seems to be the only choice in some of the application. But it can be a complementary solution for RF communications, improving the overall performances in several other scenarios. The applications can be grouped into high speed and low-speed communication types. In the high-speed applications, the primary goal of the development the higher bandwidth and the better utilization of the available bandwidth to realize higher and higher transmission bitrate.

Another classification is based on the application area; indoor and outdoor applications can be defined.

A. Li-Fi

One of the essential high-speed application envisioned for VLC is providing of Light-Fidelity or "optical Wi-Fi." In 2011, Prof. Harald Haas was the first one used the term Light Fidelity (Li-Fi) [32]. The considerable potential and the fast evolution of the Li-Fi technology contributed to the foundation of the Li-Fi Consortium in the same year [33]. Li-Fi is a high speed, bi-directional, fully connected, visible light wireless communication system and is analogous to well-known Wi-Fi. The system can enable high-speed internet connections from the ceiling lamp, because of the large available bandwidth. The limited VLC link length does not limit the scenario, because the required distance is a few meters, equivalent to the distance between the ceiling and ground in an office. This technology based on Li-Fi routers can offer multi Gb/s connections. The data, which comes from the internet, is transformed into a driving signal of the light source. According to the data, the light source will switch on and off at frequencies unperceivable by the human eye. The receiver converts the light signal into an electrical signal and next into numerical data which will be delivered to the mobile terminal.

Fig. 2. shows the basic idea of the Li-fi concept.

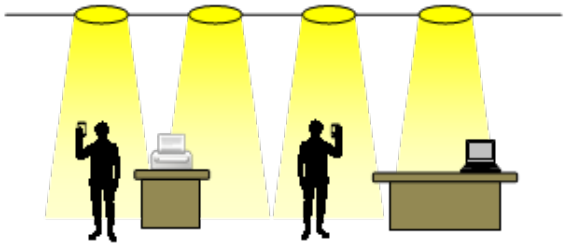


Fig.2. Li-Fi idea

There are several approaches to realize the uplink in Li-Fi system. Visible light sources could provide both up- and downlink. But the realization is not useful in all cases, because of the lack of installed VLC infrastructure for the upload. On the other hand, mainly downlink connection pushes the huge bandwidth advantage of VLC link. Typically uplink is performed using an infrared link [34], or the already existing Wi-Fi uplink [35] in integrated Wi-Fi and Li-Fi service.

The transmission of the VLC signal does not affect the already applied RF-based communications system. So the two alternative wireless technologies could be used together. It means that the two technologies can cooperate; the VLC link can take some of the load from the already full RF spectrum. One useful approach, then the VLC technology, offers high data rate downstream from the router to the mobile terminal. However, the upstream is provided by radio frequency communication from the mobile terminal to the router [35].

B. Optical-wireless communication in RF restricted areas

In areas where RF communications are limited, VLC technology can be safely used to provide wireless connections. It can be a safe alternative to radio frequency communications in hazardous environments, such as mines and petrochemical plants. It is also useful in applications where traditional WLAN communication may interfere with specialized equipment, like hospitals and aircraft passenger cabins' in-flight entertainment systems

A most studied example is the hospital environment, where electromagnetic wave sensitive areas should connect to the network. The utilization of RF communications in hospitals and health care units are restricted, especially in operating theatres and around magnetic resonance scanners. In these places, the information exchange is possible by using a VLC system, because VLC does not interfere with radio waves of the other equipment and machines.

Aviation is also a restricted area for RF communications. Radio is undesirable in passenger compartments of aircraft. LEDs are already used for illumination and can also be used instead of wires to provide media services to passengers. This approach reduces aircraft construction costs and weight. VLC can also be used in hazardous environments where there is a risk of explosions. In these areas, RF communications are restricted due to the risk. VLC can be successfully used in these areas, like mines, oil rigs, chemical plants, etc. Additionally, the communication capability is a complement to the already existing LED-based lighting systems.

On the other aspect, VLC uses light sources, which is typically not LASER, but the system must follow the safety regulations and limits. Accessible Emission Limit (AEL) is the maximum affordable emission level permitted within a particular LASER hazard class. Maximum Permissible Exposure (MPE) is the level of LASER radiation to which an unprotected person may be exposed without adverse biological changes in the eye or skin. The nominal hazard zone (NHZ) can be determined. A complex calculation may have to be performed to assess AEL, MPE, and NHZ values for certain LASERs and specific conditions. This calculation is defined in standards and regulations. However, VLC links typically apply LED-based lighting system, which does not exceed the limits, determined for LASER based system.

C. Indoor localization

In the indoor environment, VLC based localization is very convenient since the classical GPS is not able to work inside buildings [36]. The omnipresence of LEDs for illumination provides unique opportunities for indoor localization [37]. Visible light can be applied as an ID system in unusual places such as buildings and subways. The visible light ID system identifies the location, typically the room and the building. Similarly, the visible light ID system can be introduced in subways, hospitals, and airports. The system can be extended, when the ID provides the coordinates of the transmitter.

Signals transmitted by the LEDs can be used to determine the position of a person or object within a room with high precision. The requirements of such applications are different from high-speed VLC systems. One of the commonly employed methods for VLC based localization is optical tracking and imaging while the other is based on trilateration/triangulation. However, in all two purposes, it is critical that the user device can recover signals from each luminaire separately.

Almost all location determining methods are based on the a-priori knowledge of the position of communicating LED devices playing a role in the localization. Fig. 3. presents the VLC based indoor localization concept. For this, we have to know which transmitter's signal has been detected. So, multiplexing methods are required for the VLC system. Mainly three types of channel access technique can be implemented in VLC systems; time, frequency, and code division multiple access [38].

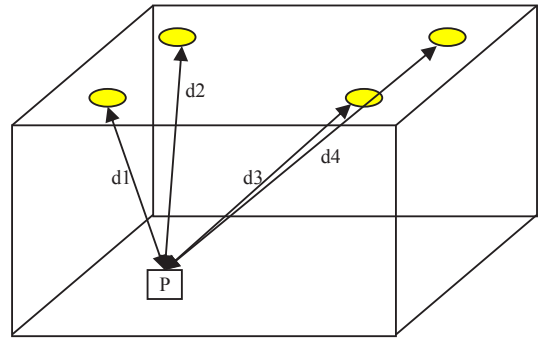


Fig.3. VLC based indoor localization

On the other hand, VLC can provide very efficient indoor localization. Centimeter accuracy can be approached by determining the received signal strength or the time of flight and by using the triangulation technique.

D. Location-aware applications

Over the years, smart advertising has become an essential marketing tool for brands to communicate with their potential customers in a place where a large number of people live. VLC could also be used to create smart areas by providing geo-localized information. This way, the location-aware information can be delivered to users' smartphones or tablets by using the indoor light. Information Displaying Signboards are often made from an array of LEDs; these signboards are installed to distribute information in airports, bus stops, and other places where the broadcasting of information is necessary. The signboard can also transmit data; this type of signboard can be applied for indications in several locations such as airports, museums, and hospitals.

The indoor public areas can be equipped with accurate positioning and localization systems based on VLC technology. Several applications may use this indoor positioning service. In theaters or cinemas, guide audience can help the visitors to their seats.

The localization information can trigger a particular audio or video guide script in a museum. The museum could use Li-Fi to provide information about exhibitions [39].

Similar to museums, retail markets could provide product information, coupon codes, and other personalized shopping experiences. It can help shoppers to find discounted items in a store or supermarket. Supermarkets are ideal advertising spaces since hundreds of consumers visit them every day and are receptive to the advertising that is offered to them. The process consists of delimiting the zones in which the owner of the supermarket wishes to deploy his advertising campaign through the positions of the LEDs suspended from the ceiling. Consumers may prefer Li-Fi in this case to Bluetooth or Wi-Fi since it allows them to remain anonymous by only receiving data.

E. Transportation

The most promising outdoor application of VLC technology is the intelligent transportation system applying the vehicle lights and the existing traffic light infrastructure. The LED headlights and taillights are commercially available in cars. The street lamps and traffic signals are also moving to LED technology. So, VLC based vehicle-to-vehicle and infrastructure-to-vehicle communications are achievable. It can manage road safety and traffic management. By using an optical wireless connection, safety messages can be transmitted from one vehicle to another one; and also from the traffic infrastructure to the approaching vehicles. Cars share data concerning their state, like velocity, location, acceleration, braking action, or their mechanical state, etc. (Fig.4.). Based on it several intelligent transport system services and applications can be realized including cooperative forward collision warning, emergency electronic brake lights, lane change warning, pre-crash sensing, stop sign movement assistant, traffic signal violation warning left turn assistant, etc. On the other hand, cars can distribute and disperse traffic information, like the place of an accident in the system. This approach can improve the safety and efficiency of the transportation system. Based on the distributed information location services and optimized alternative routes can be offered [40].

In crowded cities or on highways we have to calculate with high traffic density scenario. In this case, the ability of radio frequency communications to support intelligent vehicular communication has limited reliability because of mutual interferences. VLC is a line of sight technology, and it is not affected by interferences. So, visible light transmission can successfully support communications in high traffic density.

It is an outdoor application. So, the system must be robust to perturbations noises from the sunlight or the artificial light. The solution has also to be cost effective to make sure that high market penetration is possible.

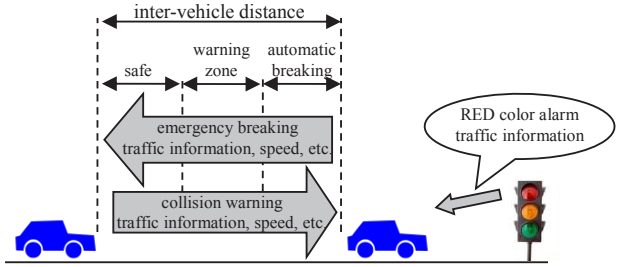


Fig.4. Intelligent transport system based on the VLC technology

F. Underwater communications

Data transmission in an unguided water environment through wireless carriers is called as underwater wireless communications. Radio-frequency wave, acoustic wave, and optical wave can be applied in this system. RF waves do not propagate well in seawater; it is not able to provide underwater communications. The bottlenecks of the acoustic transmission are the low bandwidth, high latency, and complex energy-consuming acoustic transceivers. VLC can be effectively used in this environment. In comparison to RF, underwater optical wireless communication can provide a much higher transmission bandwidth and much higher data rate. In this case, VLC can provide short-range (some meters) communications which can enable divers, underwater robots, and undersea sensors to communicate with each other or with the base station or buoy on the surface (Fig.5).

VLC technology may achieve a data rate on the order of Gbps over moderate distances of tens of meters in seawater. The transmission speed of light in water is much higher than the acoustic wave; the VLC link provides low link latency. The high transmission data rate and the low latency guarantee the realization of real-time applications like underwater video transmission. Underwater optical-wireless also has higher communication security over the acoustic and RF methods. Eavesdropping is more difficult in LoS configuration, which is implemented in VLC systems, rather than the diffused broadcasting scenario like acoustic and RF wave

communication. Finally, the optical-wireless link is much more energy efficient and cost-effective than its acoustic and RF competitors. Acoustic and RF transceivers are large, expensive, and highly energy consuming. Underwater VLC transceivers are relatively small and low-cost implementing LASER diodes or LEDs and photodiodes. [41]

Besides the advantages of marine VLC technology, there are some disadvantages or at least challenges of the approach.

The blue and green spectrum components are used for underwater communication, because of the minimum attenuation value. However, the optical signal is still degraded by the absorption, and the scattering caused multipath propagation due to the inevitable photon interactions with the water molecules and other particulate matters in the water. In an underwater environment, chlorophyll, similar issues, and colored dissolved organic materials are capable of absorbing the blue and red lights [42]. Additionally, its increase the turbidity of the water and thus shrink the propagation distance of the light. Moreover, the concentration of colored dissolved organic material will also change with ocean depth variations, and consequently, the corresponding light attenuation coefficients also vary with the water depth.

VLC communication in the marine environment is a unique situation. In this application, there is no already installed lighting infrastructure; the VLC communication link is established for the dedicate data transmission. So, blue or green LASER can be implemented instead of an LED light source. The LASER source has a narrow divergence feature, which increases the safety and the power budget. However, LASERs with a tight divergence angle require precise pointing between transmitter and receiver, an exact alignment condition is needed. This requirement will limit the performance of the system in turbulent water environments and can become a critical problem when the transmitter and the receiver are non-stationary nodes, such as an autonomous vehicle. Underwater optical links will be temporarily disconnected due to misalignment of optical transceivers. It can take place frequently, as the underwater environment is turbulent, especially in the vertical buoy-based surface-to-bottom applications. Random movements of sea surface will cause serious connectivity loss problem. Use of a LED light source with more extensive divergence may solve this problem. However, the wider divergence of the light ray comes from the LED source limits the link length.

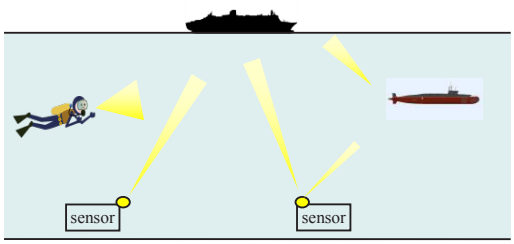


Fig.5. Underwater VLC transmission

Implementation of the systems requires reliable underwater devices. In the marine environment, the flow, the pressure, the temperature, and the salinity of seawater strongly impact the performance and lifetime of the underwater devices. The reliability of device batteries and efficiency of device power consumption are critical because solar energy cannot be exploited in the undersea environment.

V. CONCLUSION

Visible light communication has the potential to provide high-speed data communication with good security and improved energy efficiency. VLC development is motivated by the increasing demand for wireless communication technologies. The LEDs performances sustained the rapid evolution of VLC. VLC can be successfully commercialized in coming years because the interest increases from the research community, industries, and standardization.

This paper is an overview of literature covering features and applications of visible light communication. It first introduces the motivation for VLC technology development. The main advantages and disadvantages of this technology are overviewed; the challenges are presented. The main applications of VLC technology were identified.

The different modulation techniques are discussed, the optimal modulation scheme should be able to provide dimming support and minimize flickering effect while maintaining higher spectral efficiency. This part of the paper includes a review of major modulation techniques (OOK, PPM, OFDM, and CSK).

The last sections provide an overview of communication applications. High data rate indoor communication is one of the main application domains for VLC technology. It can be used for fast internet connection or a fast data broadcast over a short communication range. The scientific community has made significant efforts in this research area, which allowed VLC to obtain impressive results.

On the other hand, several outdoor applications serving more extended range are also developed by the research community. Both underwater communication and automotive area require special requirements. The main challenges in these applications are the achievement of tens of meters communication and increasing the robustness to noise. The third main area is the low bitrate approaches, mainly sensor and identification applications.

REFERENCES

[1] Murat Uysal, Carlo Capsoni, Zabih Ghassemlooy, Anthony Boucouvalas, Eszter Udvary: Optical Wireless Communications, An Emerging Technology, ISBN 978-3-319-30201-0, Springer International Publishing, 2016
[2] Zabih, Ghassemlooy; Luis, Nero Alves; Stanislav, Zvánovec; Mohammad-Ali, Khalighi: Visible Light Communications: Theory and Applications, New York, United States, Boca Raton, CRC Press, 2017
[3] Z Ghassemlooy: Visible Light Communications – A Review, MMTC Communications – Frontiers, Vol. 12, No. 3, May 2017 pp 6-13
[4] Steigerwald, D.A; Bhat, J.C.; Collins, D.; Fletcher, Robert M.; Holcomb, M.O.; Ludowise, M.J.; Martin, P.S.; Rudaz, S.L., "Illumination with solid state lighting technology," Selected Topics in Quantum Electronics, IEEE Journal of , vol.8, no.2, pp.310,320, Mar/Apr 2002.
[5] Azevedo, I.L et al.: "The Transition to Solid-State Lighting," Proceedings of the IEEE , vol.97, no.3, pp.481,510, March 2009.

[6] Hanzo, L.; Haas, H.; Imre, S.; O'Brien, D.; Rupp, M.; Gyongyosi, L., "Wireless Myths, Realities, and Futures: From 3G/4G to Optical and Quantum Wireless," Proceedings of the IEEE , vol.100, no. Special Centennial Issue, pp.1853,1888, May 13, 2012.

[7] Yiguang Wang ; Li Tao ; Xingxing Huang ; Jianyang Shi ; Nan Chi: 8-Gb/s RGBY LED-Based WDM VLC System Employing High-Order CAP Modulation and Hybrid Post Equalizer, IEEE Photonics Journal, Volume 7, Number 6, December 2015

[8] Hoa Le Minh ; Dominic O'Brien ; Grahame Faulkner ; Lubin Zeng ; Kyungwoo Lee ; Daekwang Jung ; YunJe Oh ; Eun Tae Won "100-Mb/s NRZ Visible Light Communications Using a Postequalized White LED," IEEE Photonics Technology Letters, 2009 , Volume: 21 , Issue: 15 Pages: 1063 – 1065

[9] Dong Yan1, Xurui Mao2, Sheng Xie3, Jia Cong1, Hongda Chen: Design fully integrated driver circuit for phosphorescent white Light-Emitting-Diode highspeed real-time wireless communication, accepted for publication, DOI 10.1109/JPHOT.2019.2904607, IEEE Photonics Journal

[10] Chien-Hung Yeh1 et al.: 1250 Mbit/s OOK Wireless White-Light VLC Transmission Based on Phosphor Laser Diode, IEEE Photonic journal, 2019

[11] Visible Light Communications: Theory and Applications, Chapter 5: Tamas, Cseh, Sujan, Rajbhandari, Gabor, Fekete, Eszter, Udvary, Modulation Schemes, New York, United States, CRC Press, 2017

[12] Tsonev, D.; Hyunchae Chun; Rajbhandari, S.; McKendry, J.J.D.; Video, S.; Gu, E.; Haji, M.; Watson, S.; Kelly, AE.; Faulkner, G.; Dawson, M.D.; Haas, H.; O'Brien, D., "A 3-Gb/s Single-LED OFDM-Based Wireless VLC Link Using a Gallium Nitride uLED," Photonics Technology Letters, IEEE , vol.26, no.7, pp.637,640, April1, 2014.

[13] Carlos E. Mejia ; Costas N. Georghiades: Coding for Visible Light Communication Using Color-Shift-Keying Constellations, IEEE Transactions on Communications, 2019

[14] Haiqi Zhang et al.: Gb/s Real-Time Visible Light Communication System Based on White LEDs Using T-Bridge Cascaded Pre-Equalization Circuit, IEEE Photonics Journal, 2018 , Vol.10 , Issue: 2

[15] Haoxu Li ; Jin Wang ; Xiaofeng Zhang ; Rangzhong Wu : Indoor visible light positioning combined with ellipse-based ACO-OFDM, IET Communications 2018 , Volume: 12 , Issue: 17, Page s: 2181 – 2187

[16] Tian Zhang ; Yue Zou ; Jianing Sun ; Shuang Qiao: Design of PAM-DMT-Based Hybrid Optical OFDM for Visible Light Communications, IEEE Wireless Communications Letters, 2019 , Volume: 8 , Issue: 1 Page: 265 – 268

[17] Cuiwei He ; Thomas Q. Wang ; Jean Armstrong: Performance of Optical Receivers Using Photodetectors With Different Fields of View in a MIMO ACO-OFDM System, Journal of Lightwave Technology, 2015 , Volume: 33 , Issue: 23, Page s: 4957 – 4967

[18]

[19] Xi Nan ; Ping Wang ; Lixin Guo ; Li Huang ; Zhongyu Liu: A Novel VLC Channel Model Based on Beam Steering Considering the Impact of Obstacle, IEEE Communications Letters (Early Access), 2019

[20] Kwonhyung Lee ; Hyuncheol Park ; John R. Barry: Indoor Channel Characteristics for Visible Light Communications, IEEE Communications Letters, 2011 , Volume: 15 , Issue: 2, Pages: 217 – 219

[21] Farshad Miramirkhani, Murat Uysal: Visible Light Communication Channel Modeling for Underwater Environments With Blocking and Shadowing, IEEE Access, February 14, 2018.

[22] Hasan Farahneh, Fatima Hussain, Xavier N Fernando: Performance analysis of adaptive OFDM modulation scheme in VLC vehicular communication network in a realistic noise environment, EURASIP Journal on Wireless Communications and Networking 2018

[23] Agoston Schranz, Eszter Udvary: Quantum Bit Error Rate Analysis of the Polarization-based BB84 Protocol in the Presence of Channel Errors, Photoptics 2019

[24] Ignacio Marin-Garcia ; Victor Guerra ; Patricia Chavez-Burbano ; Jose Rabadan ; Rafael Perez-Jimenez : Evaluating the risk of eavesdropping a visible light communication channel, IET Optoelectronics Year: 2018 , Volume: 12 , Issue: 6

[25] Walsh, J., Donald Wuebbles: Our Changing Climate, Climate Change Impacts in the United States: The Third National Climate Assessment, U.S. Global Change Research Program, doi:10.7930/J0KW5CXT. chapter 2, pages 19-67.

[26] James C. Lin : Current Activities on Exposure Limits for Humans in the Radio-Frequency Region, IEEE Antennas and Propagation Magazine, 2014 , Volume: 56 , Issue: 6, Pages: 256 – 258

[27] Giulio Cossu ; Raffaele Corsini ; Ernesto Ciaramella: High-Speed Bi-directional Optical Wireless System in Non-Directed Line-of-Sight Configuration, Journal of Lightwave Technology, 2014 , Volume: 32 , Issue: 10, Page s: 2035 - 2040

[28] Xun Guan, Qing Yang, Taotao Wang, Calvin Chun-Kit Chan: Phase-Aligned Physical-Layer Network Coding inVisible Light Communications, accepted paper, DOI 10.1109/JPHOT.2019.2904954, IEEE Photonics Journal

[29] Sihua Shao ; Abdallah Khreishah ; Moussa Ayyash ; Michael B. Rahaim ; Hany Elgala ; Volker Jungnickel ; Dominic Schulz ; Thomas D.C. Little ; Jonas Hilt ; Ronald Freund : Design and analysis of a visible-light-communication enhanced WiFi system, IEEE/OSA Journal of Optical Communications and Networking, Year: 2015 , Volume: 7 , Issue: 10, Pages: 960 - 973

[30] Chen Gong ; Shangbin Li ; Qian Gao ; Zhengyuan Xu, Power and Rate Optimization for Visible Light Communication System With Lighting Constraints, IEEE Transactions on Signal Processing, 2015 , Volume: 63, Issue: 16, Page s: 4245 - 4256

[31] Gábor, Fekete ; Gergely, Mészáros ; Eszter, Udvary ; Gábor, Fehér ; Tibor, Berceli: Visible light communication channel disturbances and examination of the modulation formats, International Journal of Microwave and Wireless Technologies 8 : 8 pp. 1163-1171. Paper: MRF1500 , 9 p. (2016)

[32] Harald Haas: LiFi: Conceptions, misconceptions, and opportunities, 2016 IEEE Photonics Conference (IPC), Page s: 680 - 681

[33] www.lificonsortium.org

[34] Cheng Chen ; Rui Bian ; Harald Haas: Omnidirectional Transmitter and Receiver Design for Wireless Infrared Uplink Transmission in LiFi, 2018 IEEE International Conference on Communications Workshops

[35] Aimin Tang ; Chao Xu ; Bangzhao Zhai ; Xudong Wang : Design and Implementation of an Integrated Visible Light Communication and WiFi System, 2018 IEEE 15th International Conference on Mobile Ad Hoc and Sensor Systems

[36] Cheong, Y.-K.; Ng, X.W.; Chung, W.-Y., "Hazardless Biomedical Sensing Data Transmission Using VLC," Sensors Journal, IEEE , vol.13, no.9, pp.3347,3348, Sept. 2013.

[37] Hyun-Seung Kim; Deok-Rae Kim; Se-Hoon Yang; Yong-Hwan Son; Sang-Kook Han, "An Indoor Visible Light Communication Positioning System Using an RF Carrier Allocation Technique," Journal of Lightwave Technology, vol.31, no.1, pp.134,144, Jan.1, 2013.

[38] Optical Wireless Communications, An Emerging Technology, chapter 28: Gabor Feher and Eszter Udvary, VLC Based Indoor Localization, ISBN 978-3-319-30201-0, Springer International Publishing, 2016,

[39] Minchul Kim ; Taeweon Suh: A Low-Cost Surveillance and Information System for Museum Using Visible Light Communication IEEE Sensors Journal 2019 , Volume: 19 , Issue: 4 Page s: 1533 – 1541

[40] Renata Maria Mare ; Claudio Luiz Marte ; Carlos Eduardo Cugnasca: Visible Light Communication Applied to Intelligent Transport Systems: an Overview, IEEE Latin America Transactions, 2016 , Volume: 14 , Issue: 7, Page s: 3199 - 3207

[41] Mohammed Elamassie ; Farshad Miramirkhani ; Murat Uysal: Performance Characterization of Underwater Visible Light Communication, IEEE Transactions on Communications, 2019, Volume: 67 , Issue: 1, Pages: 543 – 552

[42] Tamás, Szili ; Balázs, Matolcsy ; Gábor, Fekete: Water Pollution Investigations by Underwater Visible Light Communications Paper: Tu.P.14, 17th International Conference on Transparent Optical Networks, ICTON 2015

Eszter Udvary (M'98) received the Ph.D. degree in electrical engineering from Budapest University of Technology and Economics (BME), Budapest, Hungary, in 2009. She is currently an Associate Professor at BME, Department of Broadband Infocommunications and Electromagnetic Theory, where she leads the Optical and Microwave Telecommunication Lab. Dr. Udvary's research interests are in the broad areas of optical communications, include optical and microwave communication systems, Radio over fiber systems, optical and microwave interactions and applications of special electro-optical devices.

Performance Evaluation of Closed-loop Industrial Applications Over Imperfect Networks

Sándor Rácz, Géza Szabó and József Pető

***Abstract*—5G networks provide technology enablers targeting industrial applications. One key enabler is the Ultra Reliable Low Latency Communication (URLLC). This paper studies the performance impact of network delay on closed-loop control for industrial applications. We investigate the performance of the closed-loop control of an UR5 industrial robot arm assuming fix delay. The goal is to stress the system at the upper limit of the possible network delay. We prove that to achieve the maximum accuracy of the robot at maximum speed, URLLC is a must have.**

***Index Terms*—Industrial Application, Robot Arm, URLLC, Network Delay, Trajectory Accuracy, Measurements, Performance Evaluation**

I. INTRODUCTION

Wireless networks are continuously replacing wired networks in several areas. Mobile Broadband is one of the most successful areas. As a next step, 5G networks also provide technology enablers targeting industrial automation and control applications. One key enabler is the Ultra Reliable Low Latency Communication (URLLC). URLLC should be capable of successfully transmitting messages over radio interface within 1 ms with a 99.999% success probability and should be capable to achieve a latency of 0.5 ms on average for multiple transmissions [1].

Industrial automation and control applications require significantly different latency and reliability [2]. Least demanding applications like diagnostics and maintenance do not require latency lower than 15 ms and reliability around 99.99%. Closed-loop applications require latency between 1 ms and 15 ms and ultra high reliability. Special applications, e.g., printing machine, typically require even lower latency (<1 ms).

A well-designed Industrial-Ethernet based solution provides reliable and low latency connection [6]. A potential problem is the cable cut like events. To overcome this, aliveness of the connection is continuously monitored. In case of a connection problem, the application executes emergency action, typically, the robotic cell is stopped. For example, in ProfiNet RT [16], data frames are sent periodically (update time specifies the period) and when a predefined number of consecutive frames are not arrived in time (retry parameter specifies the threshold, typical value is 3) then the application is notified.

From controller point of view, industrial applications over wired links are designed based on the assumption of perfect communication environment, e.g., non-delayed sensing and actuation. In contrast to wired networks, providing high quality services over wireless networks is resource demanding. Wireless networks have to deal with non-negligible transmission disturbances due to e.g. interference, fading and shadowing over the radio link. Several radio mechanisms, e.g., retransmission mechanisms, active queue managements, multiconnectivity, power control, link adaptation, try to compensate disturbances, and finally the network provides high quality services.

Wired link can be replaced by wireless link without touching the control algorithm of the application when

- wireless link can guarantee the same transmission requirements as the wired link provides, or
- characteristics of wireless link fulfill the design requirements of the control algorithm.

The first case is more conservative and much more challenging to realize by a wireless network, because in several applications the underlying industrial protocol provides strict guarantees that are much higher than the application requires. In the second case, the characteristics of connection provided by wireless link are adapted to the application requirements.

The joint optimization of application control loop and wireless network can improve efficiency. If the network is informed about the current latency requirement of the control loop, then the network can more efficiently assign resources. In this way, the wireless network can serve more applications simultaneously. We address the case when neither the link nor the application are optimized jointly. This paper evaluates the performance of a robot arm control application. The application includes closed-loop control of an UR5 industrial robot arm [10] and it is connected to the robot arm through a fixed delay connection. The main focus is on the effect of the link delay on the performance of robot arm movement quality measured by specific key performance indicators (KPIs).

Our target system on which the evaluation is done is a UR5 robot arm. The UR5 is an industrial grade robot arm and has an externally accessible velocity control interface. The robot arm accepts velocity commands for each joint (servo) and publishes joint state information with 8 ms update time. Investigated KPIs are response time and precision of trajectory execution, i.e., spatial and temporal deviations from the planed trajectory.

The paper is organized as follows. Section II discusses the state-of-the-art. Section III describes the measurement setup and provides measurement results. Section IV describes the measurement scenarios. Section V shows the performed evaluations. Section VI discusses the observations. Section VII concludes the paper.

Sándor Rácz and Géza Szabó are with Ericsson Research, Budapest, Hungary (e-mail: {sandor.racz, geza.szabo}@ericsson.com

József Pető is with Budapest University of Technology and Economics, Hungary (e-mail: pjoejoejoe@gmail.com)

II. RELATED WORK

The rest of the section gives an overview of Networked Control Systems (NCSs) focusing on the introduction of wireless link.

A. Networked control systems

In networked control systems, feedback control loops are closed via communication networks [11], [12]. A NCS consists of numerous coupled subsystems, which are geographically distributed, and individual subsystems exchange information over wired or wireless networks. In [11], an overview of recent developments on NCSs is presented. Three general configurations of networked control systems, i.e., centralized, decentralized and distributed configurations are discussed. Then, challenging issues from the study and application of NCSs are outlined from three aspects: communication, computation and control. In [12], several aspects of NCS was discussed, including sampling, network-induced delays, packets dropouts, quantization errors, sampled-data control, networked control, event-triggered control, network-based filtering in continuous or discrete-time domain.

The introduction of wireless links in a NCS requires us to revisit the design aspect of the system. There are two main means to integrate a wireless link into a NCS:

- adapting the control algorithms to the properties offered by wireless link [11], [13] and
- improving the wireless network to meet the current design assumptions [9], [14] .

In [11], authors proposed a fuzzy predictive control method to mitigate the network-induced delays from sensor-to-controller and controller-to-actuator links. At each time instance, the method evaluates the network delays and the proper control law is designed based on a predictive scheme. In [13], delay compensation scheme using classical and adaptive Smith predictor was applied to wireless NCS. The Markov model was proposed to compute the estimated network delay used in the classical predictor. In the adaptive predictor, the channel delay statistics using shift registers was proposed to update the estimated delay.

In [9], they provided a low complexity RTT skew MIMO control algorithm for 5G using multiconnectivity. In [14], authors considered all control loops as network applications (i.e. keeping controllers and devices unchanged) and developed a control-aware network uplink scheduler to handle the control performance degradation caused by communication delays.

B. Wired communication

Industrial automation and control applications use diverse technology to interconnect controller and devices. Industrial-Ethernet based protocols (e.g. ProfiNet, EtherCAT) and Field-Buses (e.g. ProfiBus) are the most widespread solutions with estimated market shares of about 46% and 48%, respectively [3]. In [4], authors compare the network protocols used nowadays in industrial applications. All investigated systems show similar basic principles, which are solely implemented in different ways. Shared memory is applied and most systems require a master or a comparable management system which controls the communication. Shared memory is implemented via data distribution mechanisms that are based on a high frequency packet sending pattern. These packets have to be transmitted with strict delivery time with minimum jitter.

ProfiNet [5] distinguishes between two real-time classes with different services and target applications.

- Real-Time (RT) class: This class is suitable for applications with cycle times of 1-10 ms. Standard Ethernet components can be used to connect devices. Application, transmission and devices have their own not synchronous cycles, in this way jitter is not optimized.
- Isochronous Real-Time (IRT) class: This class is suitable for applications with cycle times of less than 1 ms. This class provides clock synchronized communication and provides jitter less than 1 μs, but needs hardware support via switch-ASIC.

C. Wireless communication

A lot of effort is put into improving radio algorithms of URLLC. In [7], authors reviewed recent advances in URLLC. In [8], authors discussed wireless channel models that are relevant for URLLC. For challenging services like URLLC, tailor-made methods are developed to achieve the strict performance requirements, e.g. [9]. The support of URLLC services comes at the cost of reduced spectral efficiency compared to mobile broadband services without latency and reliability constraints [1]. The spectral efficiency significantly depends on the provided quality, for example, URLLC providing 1 ms latency can have about 3 times lower spectral efficiency compared to URLLC providing only 10 ms latency. In networks, where the share of the URLLC traffic can be significant in the load, optimized use of URLLC can improve the network capacity.

III. EXPERIMENTAL ENVIRONMENT

In this section, we investigate the response time and the trajectory execution performance of an UR5 industrial robot arm in networked control scenarios. During measurements we used a real UR5 robot arm.

A. Hardware components

UR5 industrial robot arm is a 6-DOF, lightweight, flexible, and collaborative robot that allows to automate repetitive and dangerous tasks with payloads of up to 5 kg. The robot arm is ideal for optimizing low-weight collaborative processes, such as picking, placing and testing.

The controller of UR5 provides access to a wide range of low level functionalities. This makes the robot suitable to be included in custom networked control system. At lowest level, individual joints with brushless servo motors and harmonic drive reducers are controlled with 2.4 KHz frequency. Unfortunately, this interface is not accessible externally. The next control possibility is to command (receive) the velocity (state) of the individual joints at sampling frequency of 125 Hz and this interface is accessible externally (low level control API). The robot supports ProfiNet RT, ModBus and TCP/IP communication protocols.

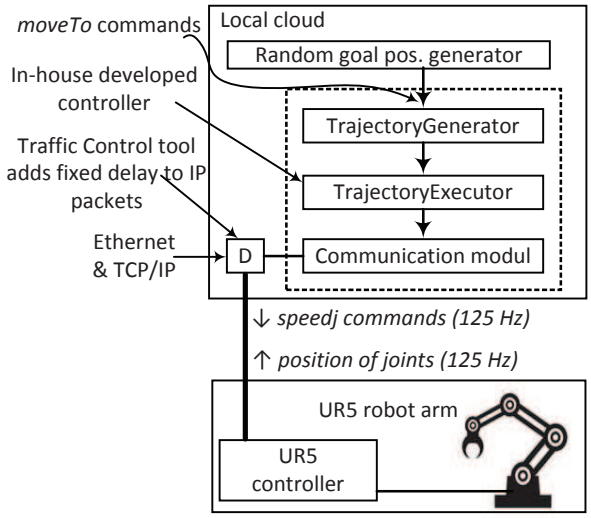


Fig. 1. Measurement setup

B. Software components

We have developed a robot arm controller from scratch based on the low level control API of the UR5. The main reason to do so is to facilitate the integration of our custom KPIs. Figure 1 shows our measurement setup. The controller runs on a Linux PC that is connected to UR5 over Ethernet. It uses TCP/IP protocol stack for communication and a 125 Hz controlling frequency. Furthermore, velocity control is applied which means that we send per joint velocity commands (*speedj*) to the robot arm every 8 ms, including rotation speed information of the 6 servo motors. Network delay that models latency aspect of URLLC link is inserted in the control loop, i.e., between the controller and UR5 robot arm, by the Traffic Control tool of Linux. We use fix delay to analyze the behavior of the system on the upper limit of the possible network delay. Note that jitter can be transformed to fix delay with a jitter-buffer.

Our controller implements trajectory generation, trajectory execution and communication modules, see Figure 1. Trajectory generator accepts *moveTo* commands. The parameters of a command specify the goal position and orientation in Cartesian space, target execution start time and maximum allowed joint velocity and acceleration. First, using inverse kinematics, the goal position and orientation are transformed into joint space. Then, a feasible path is determined from the current position of the robot to the goal position using tangent bug algorithm. Obstacles can be specified in joint space. Finally, the feasible path is sampled and cubic-spline interpolation is applied considering the specified maximum joint velocity and acceleration. The trajectory generator also supports smooth on-the-fly trajectory modification.

Trajectory executor receives trajectories. A trajectory is described by 6 splines, each spline describes individual joint position evolution in time. For each joint, a feed-forward velocity control is running with predefined update time for which default value is 8 ms. The position error is calculated from the target position coming from the spline and the current position extracted from the robot feedback. The baseline velocity is obtained from the spline by derivation and modified through a PID controller based on position error. We have tuned the parameters of the PID controller for zero network delay and we kept this setting unchanged during the investigations. The update timers of the robot and the trajectory executor are unsynchronized. This unsynchronized operation further increases the average response time with 4 ms and the standard deviation (i.e. jitter) with ~ 2.3 ms. Consequently, the average dead delay of the control loop of trajectory executor is about 18 ms and the jitter is about 3 ms. Trajectory executor also records the realized trajectory. After execution, it compares the planned and the realized trajectories and calculates KPIs.

The communication module sends the joint velocity commands to the robot. Commands are sent in clear text format, and each command message contains joint velocity values for all of the 6 joints. A velocity command message is valid until a new message received or an optionally specified timeout expired. The status feedback is encoded in binary format and has a size of about 1 Kbyte.

Network delay	Avg. response time	Stand. Dev.
0 ms	14.66 ms	1.84 ms
1 ms	14.99 ms	2.00 ms
2 ms	15.60 ms	2.50 ms
4 ms	20.61 ms	1.91 ms
8 ms	22.46 ms	1.66 ms
16 ms	30.51 ms	1.70 ms
32 ms	46.62 ms	1.81 ms
64 ms	78.50 ms	1.83 ms

TABLE I
AVERAGE AND STANDARD DEVIATION OF RESPONSE TIME FOR DIFFERENT NETWORK DELAYS

IV. MEASUREMENT SCENARIOS

A. Response time

We started with the investigation of the response time (i.e, dead delay) of the robot. We sent a (non-zero) velocity command to standstill robot and inspected the received status messages sent by the robot. The response time is the time elapsed from the command transmission to the first received status message reporting joint movements. Table I shows the mean value and the standard deviation of response times for different network delays. Without network delay, the average response time is 14.66 ms and the standard deviation is 1.84 ms. The robot checks the incoming commands periodically with 8 ms period and also sends status messages with 8 ms period. Note that the standard deviation of a continuous random variable uniformly distributed over $[0, 8)$ is 2.3. The measurements show that the internal robot operation contributes to the jitter of control-loop about 2 ms. Table also shows that the network delay additionally increases the average response time and does not significantly modify the standard deviation. This means that the quick reaction on external events needs low network delay. For example, assume that the robot moves with 1 m/sec, then e.g., 10 ms additional network delay can end in up to 1 cm additional difference.

B. Precision of trajectory execution

We evaluated three main KPIs for each trajectory to measure execution quality. Two of them measure execution precision and the third one measures the execution time. Let $\underline{p}(t)$ and $\underline{r}(t)$ denote the position functions of the planned and the realized trajectories, respectively. Positions can be defined in Cartesian space or in joint space. In Cartesian space, the 3D coordinates (i.e. x, y and z) and the orientation of the tool center point are considered. In joint space, for example, $\underline{r}(0) = \{r_1(0), r_2(0), \ldots, r_6(0)\}$ denotes the start position, where $r_i(t)$ denotes the position of i-th joint at t. Denote T_p and T_r the durations of the planned and the realized trajectories, respectively. During $T_p < t \leq T_r$, the goal position refinement is being executed by the controller. The execution is finished when the predefined goal position accuracy has been achieved or predefined refinement time limit reached. In measurements, 10 sec maximum refinement time was configured. We introduce the following KPIs:

- Spatial deviation from the planned trajectory.

$$\Gamma(t) = \min_{\tau \in [-1,1]} \left\| \underline{r}(t) - \underline{p}(t+\tau) \right\|_2, \quad t \in [0, T_r].$$

  The $\Gamma(t)$ is the minimal distance between the robot position at time t and the corresponding segment of the planned trajectory around t. For orientation,

$$\Gamma_O(t) = \min_{\tau \in [-1,1]} \arccos \left[O_r^{-1}(t) \cdot O_p(t+\tau) \right], \quad t \in [0, T_r],$$

  where $O_r(t)$ and $O_p(t)$ are unit quaternions [15] representing the realized and the planed orientations of the tool center point, respectively.
- Temporal deviation from the planned trajectory.

$$\Delta(t) = \arg \min_{\tau \in [-1,1]} \left\| \underline{r}(t) - \underline{p}(t+\tau) \right\|_2, \quad t \in [0, T_p].$$

  The $\Delta(t)$ is the time difference between minimal distance point pair at time t.
- Refinement time. $\Upsilon = T_r - T_p$ is the extra time needed to approach the goal position in the predefined spatial accuracy.

The spatial and temporal deviations describe the distance between the realized and the planed trajectory. The spatial deviation measures the distance in Cartesian space or in joint space. By temporal accuracy, we refer to the timing accuracy of trajectory execution. The temporal deviation measures how accurately the planed trajectory is followed in time. For example, assume that the robot arm exactly moves along the planned path. In this case the spatial deviation is zero. Now assume that the robot arm moves on this path with 100 ms delay, i.e. $r(t) = p(t-0.1)$. In this case the temporal deviation is -100 ms.

V. EVALUATION OF THE MEASUREMENTS

During the measurements we have executed trajectories to randomly generated goal positions and orientations. The same trajectories were executed with varying parameter settings. We used different maximum allowed join speeds (from 22.5 to 112.5 deg/sec), goal accuracy in joint space (from 0.1 to 0.001 deg), controller update times (8, 16 and 24 ms) and a wide range of network delays (RTT: 0, 1, 2, 4, 8, 16, 32 and 64 ms). The high delay values, i.e. 32 and 64 ms, are included to see extreme cases as well. Figure 2 and Figure 3 highlight measurement results.

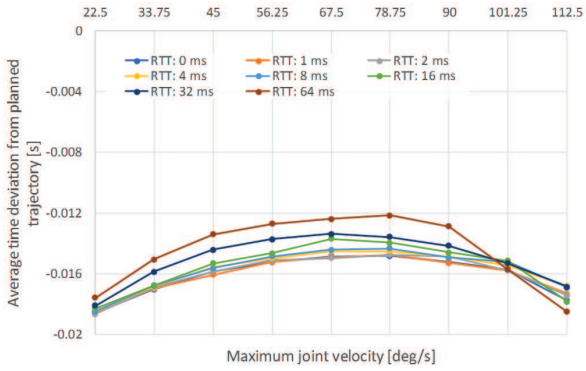
A. Affecting the temporal deviation

Figure 2(a) and Figure 2(b) show the average and the range of temporal deviation from planned trajectories for different network delays as a function of maximum allowed joint speed and using 8 ms update time and 0.1 deg accuracy. The average of temporal deviation hardly depends on the network delay, its absolute value is about 12-18 ms and the negative sign means that the robot is a little behind time in average. Note that this is approximately the dead delay of the control loop for non-delayed (i.e. RTT: 0 ms) case.

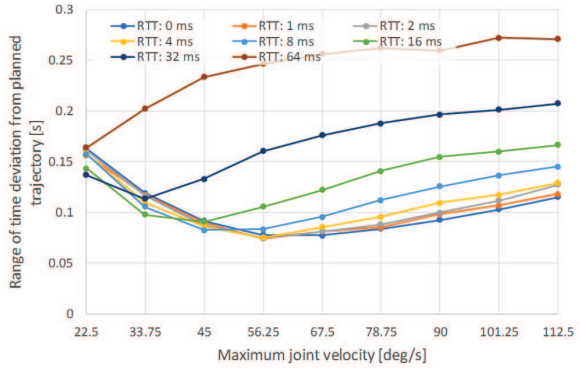
The range of temporal deviation is more sensitive to network delay. For each network delay we can observe a speed limit, e.g. for 16 ms delay it is about 45 deg/sec. If the speed is below this limit, the curve is close to the non-delayed curve. However above the limit, the range of temporal deviation curve goes above the non-delayed curve. Increased range value means that the robot is sometimes ahead of time and sometime behind time to the planed trajectory. It is also interesting that in low speed cases (e.g. 22.5 deg/sec) the range of temporal deviation is high ($\sim$ 150 ms) and hardly depends on the network delay. This can mean that in case of slow motion the high temporal deviation is probably caused by internal operation of the robot and the controller and not by the network delay. Summarizing, low network delay is required for use-cases where high temporal accuracy is crucial at high robot movement speed. For example, to avoid collision of more robot arms working close to each other.

B. Affecting the spatial deviation

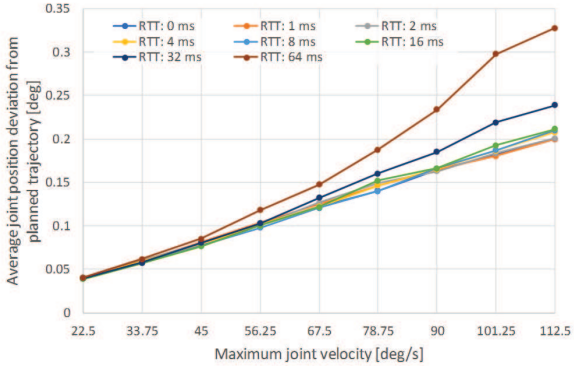
Figure 2(c) and Figure 2(d) show the average of spatial deviation from planned trajectory for different network delays as a function of maximum allowed joint speed and using 8 ms update time and 0.1 deg accuracy. In Figure 2(c), the measures are evaluated in joint space, in Figure 2(d) the measures are evaluated in Cartesian space. For higher speed, the same network delay causes higher degradation, as we expected. For network delays of 32 and 64 ms, the difference is significant. For lower network delays, the difference is relatively small. This can mean that from a certain network delay limit (in this measurement setup 32 ms) the network delay causes more intense degradation of accuracy. The two figures have similar shape. Note that, using forward kinematics formulae, joint space can be one-to-one mapped into Cartesian space. For inverse kinematics (Cartesian space to joint space transformation), a point in Cartesian space can have more than one image in the joint space. We conclude that applying lower joint speeds the spatial accuracy increases and also the controller is more tolerable to network delay. In this way, if low latency connection is available then the robot can be



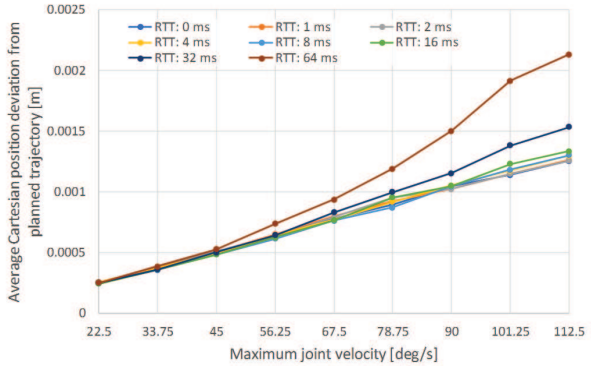
(a) Average of temporal deviation



(b) Range of temporal deviation



(c) Average of spatial deviation in joint space



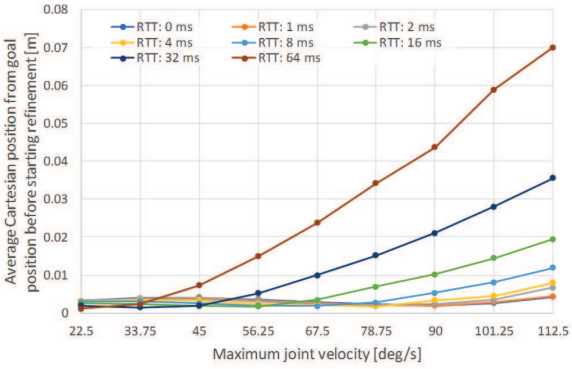
(d) Average spatial deviation in Cartesian space

Fig. 2. Statistics of temporal and spatial deviations of realized trajectory from planned trajectory for different network delays

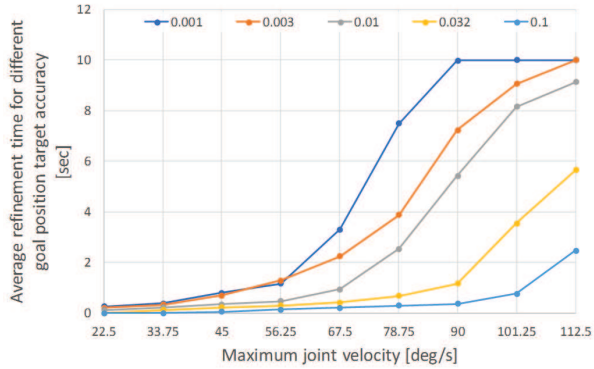
used at full speed. This also means that if only higher latency connection is available then using lower robot speed allows achieving the same spatial accuracy.

C. Affecting the refinement time

Figure 3 shows results for refinement time. Figure 3(a) shows the spatial distance at time T_p, i.e., when the planned trajectory ends and goal refinement starts. As we expected, this KPI has similar figure as average spatial deviation values. Figure 3(b) shows how refinement times depend on the required spatial accuracy (8 ms update time and 16 ms network delay). Refinement time is higher for stricter accuracy requirement and for higher robot speed. There are also cases when the required accuracy cannot be achieved within predefined time limit, e.g., 0.001 deg accuracy and > 90 deg/sec speed. This means that a deadline on execution time leads to requirement on maximum tolerable network delay. In contrary, using lower robot speed with the same spatial accuracy is achievable over a connection with higher latency. The cost is the increased execution time. Consequently, choosing proper required accuracy can improve execution time. In a robotic cell, the cyclic time is an important KPI. The cyclic time can be improved by reducing refinement time by specifying lower accuracy for cases where spatial accuracy is not crucial.



(a) Average spatial distance at the start of goal position refinement



(b) Average refinement time for different required goal position accuracy in joint space; accuracy requirements: from 0.001 deg to 0.1 deg

Fig. 3. Statistics of goal refinement phase

Max joint speed	Delay	8 ms tick	16 ms tick	24 ms tick
22.5 deg/sec	16 ms	0.116	0.116	0.115
	64 ms	0.143	0.186	0.141
45 deg/sec	16 ms	0.229	0.232	0.255
	64 ms	0.622	0.710	0.675
67.5 deg/sec	16 ms	0.340	0.374	0.492
	64 ms	1.464	1.571	1.568
90 deg/s	16 ms	0.477	0.608	0.808
	64 ms	2.531	2.660	2.793
112.5 deg/sec	16 ms	0.801	1.029	1.287
	64 ms	3.694	4.144	4.100

TABLE II
MAXIMUM SPATIAL DEVIATION IN JOINT SPACE [DEG] FROM PLANNED TRAJECTORY FOR DIFFERENT CONTROLLER UPDATE TIME (TICK), NETWORK DELAY AND MAXIMUM JOINT SPEED.

D. Experimenting with the update time of the controller

In the final measurement, we investigated the effect of update time of controller on the accuracy. In Table II, maximum spatial deviation in joint space is shown. For each trajectory, the maximum spatial deviation is calculated (i.e. $max_{t\in[0,T_B]}\Gamma(t)$) and averaged over executed trajectories. We observed that to utilize the advantage of lower update time requires low network delay as well. For example, in 64 ms network delay cases, using the lowest (i.e. 8 ms) update time has no significant effect on the performance. However, for lower network delay cases (e.g. 16 ms), lower update time leads to significant gain. This means that systems using low update time require strict latency requirements from the wireless link. Providing low latency connection for a system with high update time has no performance advantage.

VI. DISCUSSION ON THE OBSERVATIONS

This section summarizes and discusses observations and also suggests a method to handle loss and jitter.

A. Requirements on the network

In general, measurement results have shown that the network delay lower than 4 ms has no significant performance impact. This is because (a) the internal operation of the robot ends in about 2 ms standard deviation in response time, most probably, due to the internal sampling used in the robot and (b) the ticks of the robot and the controller are unsynchronized. The impact of network delay lower than 4 ms is masked by the background "noise" of measurement setup. The detailed analysis of 0-4 ms network delay range requires more sophisticated measurement apparatus, e.g., the robot and the controller should be synchronized, otherwise the randomness introduced by unsynchronized update times dominates the behavior or a robot arm with lower update time (e.g., < 1 ms) should be used.

The task of the robot arm can put requirements on the network delay:

- For tasks where robot arm should react on external events, low network delay is desired, because the network delay between robot and controller directly increases the reaction time.
- For tasks where time consuming goal refinement is not tolerable, low network delay should be provisioned. The deadline on trajectory execution time leads to a requirement on the maximum tolerable network delay. In general, higher network delay makes the refinement time longer and in this way increases the total trajectory execution time.
- Some tasks require accurate movement along the path, e.g. welding, and not only at the goal position. Another example is the collaboration of more robot arms where the precise and synchronized movements are crucial. For these tasks also low network delay is desired.

The internal mechanisms of robot arm can also put requirements on the network delay. In general, a low update time system requires lower network delay. The control of a robot arm with e.g. 20 ms update time, probably tolerates higher network delay than a more precise and faster robot arm with e.g. 1 ms update time. In addition to this, providing low latency connection for a system with relatively high update time has limited performance advantage.

Performance requirements of trajectory execution can also put requirements on the network delay. Faster robot movements require lower network delay for accurate movement. In other side, if only higher latency connection is available then using lower robot speed can compensate increased network delay for some extent.

Performance optimization can also give guidelines for required network delay. Choosing proper required accuracy can improve execution time. For example, if less accurate movement is enough, then relaxed accuracy can shorten refinement time.

B. Handling jitter, delay and packet loss

In the measurements, jitter and packet loss were not considered. We assumed negligible jitter and no packet loss was introduced by the network.

When the *jitter* is relatively small compared to the latency of the connection, then jitter buffer like methods can be used to transform jitter into extra delay. During the end-to-end delay budget calculation this extra delay should be taken into account. This method requires packet buffering capability.

Delayed or lost packets that were not arrived in time can end in performance degradation. The best solution is to minimize the occurrence of these events and to avoid bursty occurrence of them. One of the main goals of URLLC is to provide reliable connection and fulfill these requirements. In some extent, delayed or lost packets can also be handled at higher layers in the controller and at the device side. All correction methods reduce the accuracy of movements and can efficiently be used only for a limited time period.

In case of delayed or lost *status messages*, action should be taken at the controller side of the control loop. In case of trajectory execution, the controller uses joint positions

from the status messages. The missing joint position values can efficiently be extrapolated, because trajectory generators intentionally generate smooth trajectories to reduce the load of the servos. Practically, the missing joint position value is extrapolated from the historical values of joint positions and from the remaining part of the trajectory. The position error caused by extrapolated values will be corrected by the PID control when the controller receives again correct status messages.

In case of delayed or lost *command messages*, action should be taken at the both sides of the control loop. At the *controller side*, the controller needs to be informed about the unsuccessful command transmission to keep itself up-to-date. A potential solution is that the wireless network informs controller about transmission status of down-link packets. When radio interface failed or predicted to fail to transmit a packet in time (e.g. radio related problems or congestion), then wireless network notifies the controller about this event. Relying on this information the controller updates its internal state and tries to avoid overreaction.

VII. CONCLUSION

We investigated the performance of the closed-loop control of an UR5 industrial robot arm at varying network characteristics. We run trajectories and measured the accuracy of realized trajectories as the function of network delay and robot movement speed. We introduced KPIs to evaluate the temporal and spatial accuracy of the realized trajectories. We observed that to achieve the maximum accuracy of the robot at maximum speed, there is a need for low latency communication. However, at lower speed or at relaxed accuracy, higher network latency is still tolerable. We also observed that, providing much lower latency than the update time of the robot has only moderate performance gain. Finally, we suggested a method to handle loss and jitter of robot control packets.

REFERENCES

[1] J. Sachs, G. Wikstrom, T. Dudda, R. Baldemair and K. Kittichokechai, "5G Radio Network Design for Ultra-Reliable Low-Latency Communication," in IEEE Network, vol. 32, no. 2, pp. 24-31, March-April 2018.

[2] S. A. Ashraf, I. Aktas, E. Eriksson, K. W. Helmersson and J. Ansari, "Ultra-reliable and low-latency communication for wireless factory automation: From LTE to 5G," 2016 IEEE 21st International Conference on Emerging Technologies and Factory Automation (ETFA), Berlin, 2016, pp. 1-8.

[3] Automation Inside (March) 2017 [Online]. Available: http://www.automationinside.com/2017/03/industrial-network-market-shares-2017.html

[4] P. Danielis, J. Skodzik, V. Altmann, E. B. Schweissguth, F. Golatowski, D. Timmermann and J. Schacht, "Survey on real-time communication via ethernet in industrial automation environments," Proceedings of the 2014 IEEE Emerging Technology and Factory Automation (ETFA), 2014, pages 1-8.

[5] PROFIBUS and PROFINET International. (April 2019) [Online]. Available: https://www.profibus.com/

[6] S. Nsaibi, L. Leurs and H. D. Schotten, "Formal and simulation-based timing analysis of Industrial-Ethernet sercos III over TSN," 2017 IEEE/ACM 21st International Symposium on Distributed Simulation and Real Time Applications (DS-RT), Rome, 2017, pp. 1-8.

[7] Mehdi Bennis, Merouane Debbah and H. Vincent Poor, "Ultra-Reliable and Low-Latency Wireless Communication: Tail, Risk and Scale," CoRR abs/1801.01270 (2018)

[8] Patrick C. F. Eggers, Marko Angjelichinoski and Petar Popovski, "Wireless Channel Modeling Perspectives for Ultra-Reliable Communications," CoRR abs/1705.01725 (2017)

[9] R. A. Delgado, K. Lau, R. H. Middleton and T. Wigren, "Networked Delay Control for 5G Wireless Machine-Type Communications Using Multiconnectivity," in IEEE Transactions on Control Systems Technology, 2018, Early Access.

[10] https://www.universal-robots.com/products/ur5-robot/

[11] N. Vafamand, M. H. Khooban, T. Dragicevic and F. Blaabjerg, "Networked Fuzzy Predictive Control of Power Buffers for Dynamic Stabilization of DC Microgrids," in IEEE Transactions on Industrial Electronics, doi: 10.1109/TIE.2018.2826485

[12] X. M. Zhang, Q. L. Han and X. Yu, "Survey on Recent Advances in Networked Control Systems," in IEEE Transactions on Industrial Informatics, vol. 12, no. 5, pp. 1740-1752, Oct. 2016.

[13] Mahmoud Gamal, Nayera Sadek, Mohamed R.M. Rizk and Ahmed K. Abou-elSaoud, "Delay compensation using Smith predictor for wireless network control system," Alexandria Engineering Journal, Volume 55, Issue 2, 2016, Pages 1421-1428.

[14] Q. Liu, S. Zoppi, G. Tan, W. Kellerer and E. Steinbach, "Quality-of-control-driven uplink scheduling for networked control systems running over 5G communication networks," 2017 IEEE International Symposium on Haptic, Audio and Visual Environments and Games (HAVE), Abu Dhabi, 2017, pp. 1-6.

[15] Huynh, D.Q., "Metrics for 3D Rotations: Comparison and Analysis," J Math Imaging Vis (2009) 35: 155. https://doi.org/10.1007/s10851-009-0161-2

[16] PROFINET Real-Time Protocol (PN-RT) https://wiki.wireshark.org/PROFINET/RT

Sándor Rácz received his MSc and PhD in electrical engineering from the Budapest University of Technology and Economics (BME), at the Department of Telecommunications and Media Informatics (TMIT) in 1997 and 2004 respectively. Since 2000, he has been a research fellow at the Ericsson Traffic Analysis and Network Performance Laboratory (Traffic Lab) in Budapest. His research interests include performance modelling and analysis of telecommunication systems. He published several patents, as well as conference and journal papers, for which he received more than 850 independent citations.

Géza Szabó is working as a Senior Researcher in the Artificial Intelligence research area in Ericsson Research and taking part in designing and implementing demonstrations for external events within our robotics team e.g., Mobile World Congress, Hannover Messe.

József Pető received his MSc. degree in Computer Engineering from the Budapest University of Technology and Economics in 2018. He is currently working toward his Ph.D. degree. His current areas of interest and research include cloud robotics, digital twin, robot simulation, machine learning applications in robotics.

Wireless Authentication Solution and TTCN-3 based Test Framework for ISO-15118 Wireless V2G Communication

Zoltán Jakó, *Member, IEEE*, Ádám Knapp, *Member, IEEE* and Nadim El Sayed

***Abstract*— Vehicle to grid (V2G) communication for electric vehicles and their charging points is already well established by the ISO 15118 standard. The standard allows vehicles to communicate with the charging station using the power cable, i.e. a wired link, but it is improved to enable wireless (WLAN) links as well. This paper aims to provide an implementation that accomplishes a wireless authentication solution (WAS). With that the electric vehicles can establish V2G connection when approaching the charging pool, then identify and authenticate the driver and/or the vehicle. Furthermore, the paper presents a TTCN-3 based validation and verification (V&V) framework in order to test the conformance of the prototype implementation against the standard.**

***Index Terms*—Vehicle-to-Grid, ISO 15118, wireless charging, Electric Vehicle, ITS, TTCN-3**

I. INTRODUCTION

The proportion of Battery Electric Vehicles (BEV) and Plug-In Hybrid Electric Vehicles (PHEV), against conventional vehicles with internal combustion engine, is growing remarkably in developed countries. Led by the USA, the European Union and Japan the BEV and PHEV market is rapidly growing [1]. To serve this increased demand, massive charge point deployment is required. Nevertheless, due to business issues (e.g. billing) and grid limitations, smart charging is also a mandatory requirement to overcome the issues caused by mass electric vehicle (EV) recharging. For the sake of convenience hereafter the collection term EV for both battery electric vehicles and PHEVs is used.

The communication between EVs is an extensively researched topic and it is becoming an essential part of the C-ITS (Cooperative Intelligent transportation system) environment. The bi-directional communication between the vehicle and the charging point (and the grid infrastructure behind it) is referred to as vehicle-to-grid (V2G), thus V2G provides a communication interface for bi-directional charging (or discharging) of EVs. The EV charging station is the so-called EVSE (Electric Vehicle Supply Equipment). Inside the EV there is a module responsible for the V2G communication. This module is referred to as Electric Vehicle Communication Controller (EVCC), while in the case of EVSE the literature uses the term Supply Equipment Communication Controller (SECC). The EV is capable of communicating with the charging point using its EVCC. The message exchange between the EV and the EVSE is standardized by ISO/IEC (International Organization for Standardization/ International Electrotechnical Commission) in the series of 15118 (e.g. [2] – [7]). As the communication parts of this generic equipment are the EVCC and SECC, ISO 15118 describes the communication between these components. ISO 15118 is the enabler of vehicle-to-grid applications.

The main challenge of any standardized technology is conformance and interoperability. Conformance testing checks a specific product (or maybe a part of a product) for compliance to requirements given in a base standard. A definition of interoperability testing is the "ability" of two or more systems (or components) to exchange and use information and execute successful procedures/sessions. The aim of interoperability testing is not restricted to demonstrating that products (from different manufacturers) can work together: it also shows that these products can work together using a specific protocol. Multi-vendor compatibility is crucial for the success of V2G technology.

The contribution of this manuscript is given as follows:

1. Introduce a prototype SECC implementation, which uses wireless (WLAN-based) communication to handle a V2G session with the EVCC. A wireless authentication solution (WAS) is presented that allows and handles the V2G communication and the identification of the EV via wireless links.
2. Provide a validation and verification (V&V) tool to test the V2G conformance of the implemented prototype against the base standard given in [3].

It is important to highlight the fact that V2G was originally planned to be used in a wired manner (i.e. using the charging cable with power line communication). However, wireless communication recently gained higher attention, even in the standardization process [8], [9]. Wireless communication between EV and EVSE is based on WLAN (802.11n or Wi-Fi). Hereafter the term wireless link is used, noting that it actually denotes WLAN in the context of this manuscript. To be more precise, the ISO 15118 foresees the option of wireless authentication especially in the draft of its sixth part [7] (more details are given in Section II.B). The wireless interface allows

This work is a part of the project NeMo - Hyper-Network for electro-Mobility that received funding from the European Union Horizon 2020 research & innovation program under grant agreement no 713794. Content reflects only the authors' view and European Commission is not responsible for any use that may be made of the information it contains.

Zoltán Jakó is with the Broadbit Hungary Kft., 1023, Ürömi utca 40, Budapest, Hungary (e-mail: zoltan.jako@broadbit.net).

Ádám Knapp is with the Broadbit Hungary Kft., 1023, Ürömi utca 40, Budapest, Hungary (e-mail: adam.knapp@broadbit.net).

Nadim El Sayed is with the DAI-Labor, Technische Universität Berlin, Berlin (TUB), Berlin, Germany (e-mail: nadim.elsayed@dai-labor.de).

the EV driver to start the V2G communication (and/or use optional value-added services) before parking. If the charge point is reserved, then the EV driver may be notified via wireless interface before parking. With wired communication this is only possible after parking and plugging the EV.

The conformance testing framework is based on script language used for testing purposes, the so-called TTCN-3 (Testing and Test Control Notation version 3) [10]. V2G has massive literature background related to security issues and performance tests. However, the conformance testing of the V2G protocol itself is less discussed. On the other hand, this is also a relevant issue, which enables the spreading of V2G technology worldwide.

A. Related Works

The first significant V2G related test paper was presented by Project eNterop [11]. They had created a conformance testing setup that is for black box testing of connected Systems Under Test (SUT) [12]. They define conformance tests, which can be fully automated. Furthermore they applied TTCN-3 scripts and later this test setup was used in ISO 15118-4 [5]. Shin *et al.,* in [13] provides a test system for EVSE in accordance with relevant standards, including ISO-15118-2,3 ([3], [4]) IEC-61851, IEC 61850-90-8 and HPGP (HomePlug Green PHY – Power line communication).

Compared to these related works, our conformance testing framework differs in two aspects. First, our conformance testing framework is using Ericsson's Titan TTCN-3 complier [14], which is now open source. Therefore, there is no need to buy expensive software to compile TTCN-3 scripts. Secondly, in this manuscript the focus is on the wireless (WLAN based) communication between the tested system and the conformance test tool. This is a completely new paradigm, therefore the standardization process has just began [6], [7].

The manuscript is organized as follows. Section II gives a brief introduction to the series of ISO 15118. Section III introduces the proposed WAS, meanwhile Section IV presents the conformance testing framework. Finally, Section V gives concluding remarks and concludes the paper.

II. STANDARDS OF ISO 15118

The series of ISO 15118 standard currently contains nine parts. Each part is responsible for a small piece of the field of V2G. In this section, a brief overview of this standard family is given.

ISO 15118-1 has the title „General information and use-case definition". This document collects the use cases and overall goals of the standard itself [2].

The second part [3] is the most important from all for us, since it defines the technical specifications of all application layer messages and their respective parameters exchanged between the EV and the EVSE.

The (wired) physical and data link layer requirements are given in ISO-15118-3 [4]. Power line communication as defined in the HomePlug Green PHY specification is applied to encode digital signals onto the Control Pilot (CP) pin, which is part of the charging cable. These layers establish the Higher-Level Communication (HLC) outlined in ISO 15118-2. This third part also concerns the interaction with another standard called IEC 61851. This specifies analogue signals that encode the available amperage at a charging station. ISO 15118 builds upon this analogue and mainly safety-related IEC standard and enhances the charging process with digital higher-level communication.

Part 4 [5] is also important from the perspective of this manuscript. This part contains the conformance tests (TTCN-3 scripts) for the requirements specified in ISO 15118-2. Note that part 4 also contains lower layer test cases related to the wired link that are not considered in the present prototype system.

Part 5 is currently under preparation. When it is finalized, it will contain the conformance tests for the physical interface and its requirements defined in ISO 15118-3 [4].

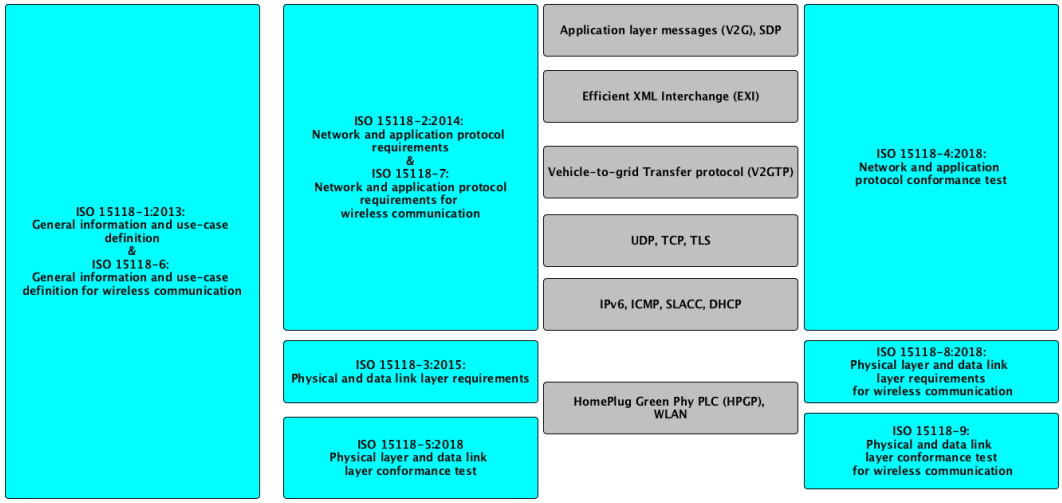


Fig. 1: Relationship between ISO 15118 parts

ISO 15118-6 [7] collects the general information and use-case definition for wireless communication, similarly to ISO 15118-1. It is foreseen that Part 1 and 6 will be merged into one document in the near future and both wired and wireless communication use cases will be available in the next version of ISO 15118-1.

The network and application protocol requirements for wireless communication will be presented in Part 7, however this document does not yet exist. It is also foreseen that Part 2 and 7 will be merged into one document.

Part 8 [6] is similar to Part 3, the big difference between them is that it contains the physical layer and data link layer requirements for wireless communication. The first version of 15118-8 was published in the first quarter of 2018.

Finally, Part 9 shall contain the conformance tests for wireless charging. On the other hand, Part 9 is under development, therefore, there is no document available yet.

An overview of ISO 15118 and the relationship between the parts are illustrated in Fig. 1.

A. V2G protocol stack (PLC)

The protocol stack of the V2G is presented in this subsection. The whole protocol stack is lavishly detailed in ISO-15118-2 [3], therefore we just give a brief presentation in this subsection. After the plug of the EVSE is connected to the EV an IPv6 address is assigned to the EV (the physical- and MAC layer link is established). The IPv6 address is assigned to the EV by DHCPv6 (Dynamic Host Configuration Protocol) and Stateless auto-configuration (SLAAC). Note that SLAAC is mandatory, but DHCPv6 is optional according to the standard.

Subsequently, the EV shall send a SECC Discovery Request message as UDP multicast over IPv6. The SECC receives and replies to the request with a response message containing the link-local IPv6 address of the EVSE. This message exchange is the so-called SECC Discovery Protocol (SDP). Afterwards the HLC can start. HLC is the bidirectional digital communication that uses the protocol and messages specified in ISO 15118-2 and ISO 15118-3 (or 15118-7). HLC includes the Protocol Handshake using the Vehicle to Grid Transfer Protocol (V2GTP), over the Efficient XML Interchange (EXI) format, and the V2G messages (e.g. Session Setup Request message). HLC allows, among other things, to negotiate the charging parameters and to authenticate and authorize the EV and the user, utilizing more secure cryptographic certificates in the plug and charge case.

The EV-EVSE can send or receive V2G application layer messages. On top, the possible message set is selected based on the usage. There are common V2G application layer messages and there are some sets related to the charging type (e.g. AC, DC or inductive charging, etc.). The V2G messages are described in the format of XML (Extensible Markup Language). A plain XML message contains significant overhead and unnecessary information (unnecessary regarding the EVCC or SECC part). Therefore, to reduce the size of the XML message it is encoded into EXI format. The resulting data is encapsulated into the V2GTP, which is encrypted using the TLS protocol, and transmitted using the general TCP/IP protocol suite to the EV or EVSE. The standard defines a couple of possible data links and physical layers as well [3], [6].

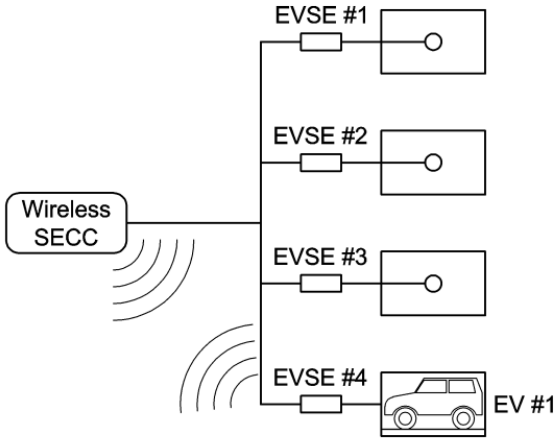


Fig. 2: Wireless communication between SECC and EV(s) as described in ISO 15118-6 [7]

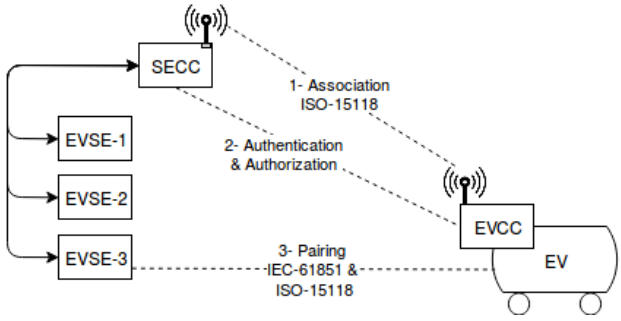


Fig. 3: Association and Pairing sequence

B. Wireless V2G

Unlike the PLC case, which may apply only to conductive charging, the wireless communication allows the support of more use cases such as the static inductive charging or the dynamic Wireless Power Transfer (WPT). The wireless parts of the ISO 15118 (parts 6, 7 and 8) base their considerations on three entities already defined in the ISO 15118-1. These are the following: EVSE, EVCC and SECC.

Unlike in the wired case (plug and charge), where the communication is rather point-to-point, the wireless communication is point to multipoint, which creates several challenges for the communication integrity, confidentiality and authenticity [15]. Thus, the ISO 15118 foresees an additional pairing mechanism to make sure that the EV, which is (wirelessly) communicating to the EVSE is in fact the exact one plugged at the Charge point or driving over the coil in case of WPT.

The main difference between the wireless V2G and the PLC V2G is that in the PLC case, the communication starts when the car is plugged, and the communication partner (EVCC and SECC) are unambiguously identifiable. Furthermore, the SECC knows exactly at which EVSE the EV is plugged. Where as in the wireless V2G, this is not true, and the wireless V2G protocol needs to define the necessary means to ensure unambiguity, confidentiality, mutual integrity and authenticity.

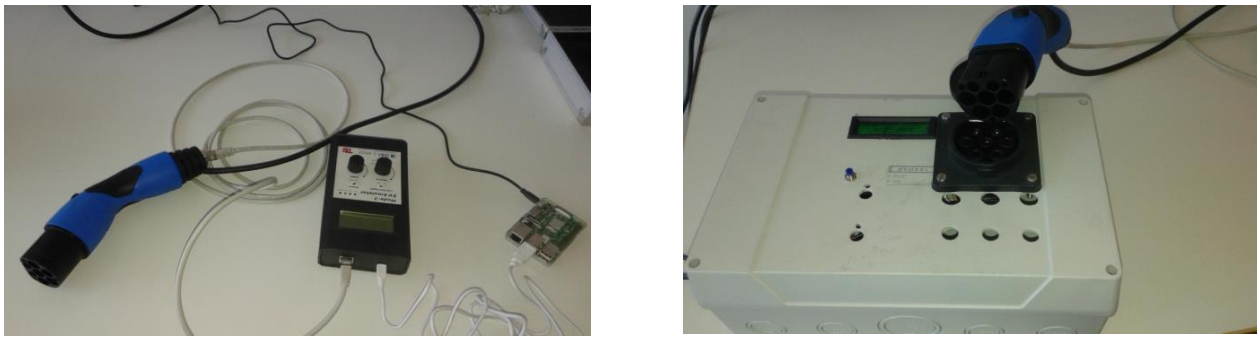


Fig. 4: Implemented EVCC (left) and EVSE (right)

The wireless communication between the EV and the EVSE is depicted in Fig. 2 and in Fig. 3. Each EVSE can be connected to one SECC only and the EVCC is able to communicate with the SECC over the wireless link. Furthermore, the association is defined as the process of establishment of wireless communication between SECC and EVCC. The Discovery is the phase in which EV obtains a list of available SECCs in its wireless communication range. This is handled by the SDP, similarly to wire environment. On the other hand, pairing is the process by which a vehicle is correlated with the unique EVSE at which it is located and from which the power will be transferred either through a cable or through wireless technology.

Pairing is done after the SDP and association phases. EVCC asks SECC the authorization to start a Pairing sequence. After a positive answer from SECC Pairing starts. EV starts the sequence of B-State, C-State, B-State toggle (referring to the different vehicle states from IEC 61851). The EVSE that detects the sequence of toggles informs SECC of the pairing toggles detection. SECC informs EVCC of the correct toggles reception. Depending on whether the location detected is convenient or not, SECC may decide to ask EVCC to change location. The implementation details are described in [16].

Once SECC Discovery, Association, and Pairing are done, the V2G application layer communication (i.e. HLC) can start. Note that this approach does not depend on the high precision localization (e.g. GPS) to determine the proximity of the EV to a certain EVSE. The following Section describes the prototypical implementation of our wireless authentication solution.

III. PROPOSED WIRELESS AUTHENTICATION SOLUTION (WAS)

This prototype implements the wireless communication for the conductive charging case, yet most of the components are applicable to inductive charging. This is especially true for the EVCC – SECC communication (SDP and Association), and the high level communication (V2G application layer message exchange). Merely the EVSE and the EV parts have to be adapted to implement the correspondent standards for wireless power transfer, which affects the pairing (and fine positioning) part of the implementation. Our implementation consists of three entities: EVCC, SECC and EVSE. Each of these entities is composed in its turn of different components that provide different functionalities. These components can interact through interfaces. The functionalities of the prototype implementation of the ISO 15118 based wireless authentication cover all the layers of the ISO/OSI stack. It is important to highlight the fact that this implementation does not focus on the charge process (energy flow) itself, only on the communication (V2G) part. Therefore, there is no need to implement all the V2G message set (e.g. messages responsible for metering data exchange).

A. EVCC

The EVCC implementation refers to the conductive charging that uses the IEC 62196 Type-2 Connector Plug (illustrated in Fig. 4). The EVCC consists of the following components: EV-Emulator and EV-Controller.

The *EV-Emulator* is built up by off the shelf microcontroller (Atmel ATMEGA16p) and circuit elements for implementing the IEC 61851 functionalities necessary for the pairing, which is achieved by doing some toggling pattern on the wire, based on ISO 15118-3 [4]. Furthermore, the EV-Emulator implements a UART interface to the EV-Controller. Meanwhile the *EV-Controller* is a Raspberry Pi 3 device equipped with a WLAN module (802.11n) for discovering the different SECCs in the neighbourhood and reading out their Vendor Specific Elements (VSE) on the ISO-Layer 2 containing their EVSEID according to ISO 15118-8 [6]. The EV-controller associates with a SECC using IPv6 Stateless auto-configuration and implements a SECC Discovery Protocol (SDP) Client to get the SECC settings and endpoint parameters over UDP-Multicast. Using these parameters, the EV-controller performs a TLS handshake with the SECC by verifying the Root-V2G

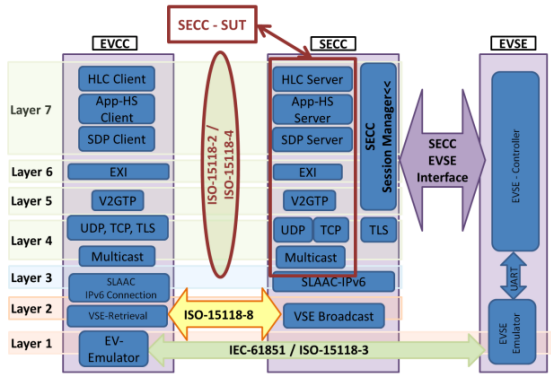


Fig. 5: WAS and SECC SUT

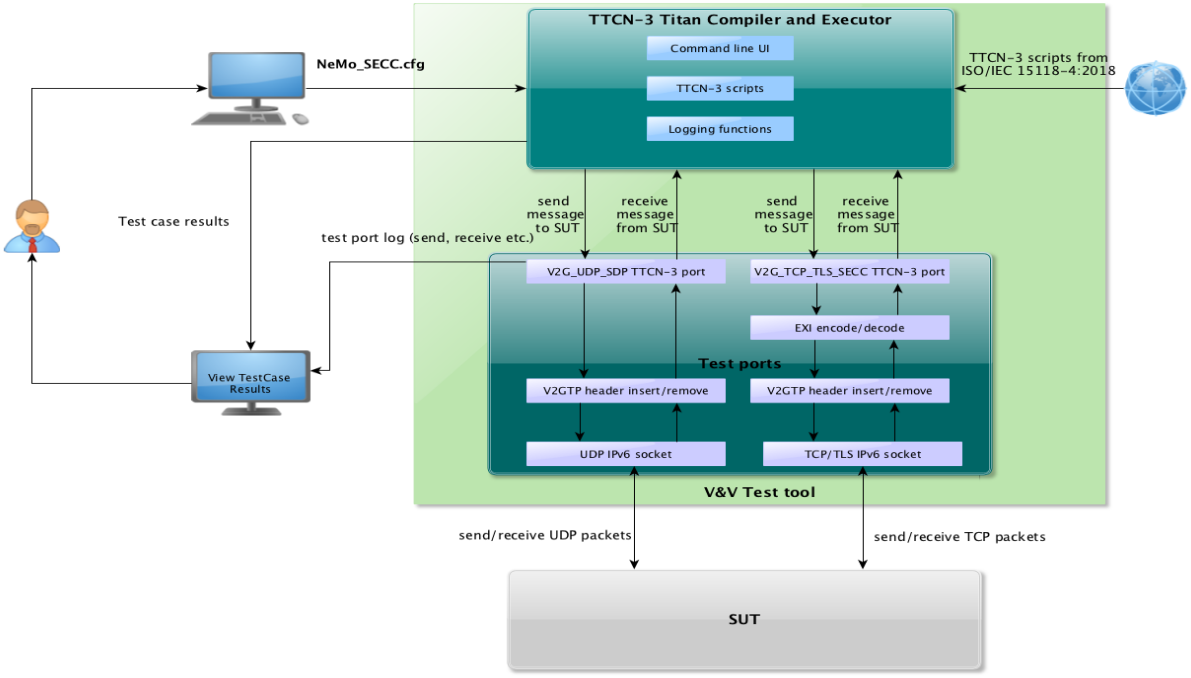


Fig. 6: The architecture of the Validation and Verification (V&V) test tool including the SUT

certificate, then its state machines start the HLC-Communication by implementing the client side of the ISO 15118-2 [3]. This is an EXI encoded communication, encapsulated in the Vehicle to Grid Transfer Protocol. When in pairing state, the EV-Controller sends the toggles through the EV-Emulator over the UART interface.

B. *EVSE*

The EVSE consists of the following components: EVSE-Emulator and Charge Point Manager (CPM).
The EVSE-Emulator is an OpenEVSE hardware [17][15], connected to an IEC 62196 Socket (see Fig. 4). Its firmware allows us to control and access the lower level functionalities of the EVSE according to the IEC 61851, which makes it the counter part of the EV-Emulator for reading the toggling on the wire. It implements a UART interface to the Charge Point Manager. The Charge Point Manager (CPM) is a Raspberry Pi 3 device connected with the EVSE-Emulator over USB, and implements the higher level messages of the pairing process. The CPM reads the toggles on the wire from the EVSE-Emulator over the UART, and communicates them to the SECC, that in his turn tracks which EVSE has detected the respective EV.

C. *SECC*

With regard to wireless V2G communication, the SECC sub component is the most important part of our implementation.
The SECC consists of the following components: SECC-Controller and Charging Session Manager (CSM). Again, SECC-Controller is a Raspberry Pi 3 device equipped with a WLAN module, which operates in the Access Point (AP) Mode and spans an Automatic Wireless Charging (AWC) network, and includes its EVSEID in its VSE. It listens on the Multicast group and implements the SDP Server. It also implements the Server side of the HLC Communication. In coordination with the Charging Session Manager (CSM) denotes the central logical component of the SECC where the other components are linked together. It tracks the different charging sessions of the different EVs and their respective EVSE, including their state. It interfaces with the SECC-Controller for the EV communication and with the EVSE-Controller for the charge point information retrieval and control.

IV. TESTING WIRELESS AUTHENTICATION SOLUTION WITH TTCN-3

In this section, a V&V testing framework is introduced for V2G communication. As mentioned in Section II regardless of the physical layer, the network and application requirements are common for any ISO 15118-based implementation. This is why the focus of the testing framework developed here is mainly based on the standard ISO 15118-2 (more precisely ISO/IEC 15118-2-ED2) and its respective conformance tests specified in the ISO/IEC 15118-4 [5]. This conformance testing framework is applied on the SECC part of the WAS presented in Section III.

A. *Test bed*

Conformance tests specify the testing of capabilities and behaviours of a System Under Test (SUT), as well as check what is observed against the conformance requirements specified in ISO/IEC 15118-2 [3] and against what the supplier states the SUT implementation's capabilities are. In this case, SUT is a software that implements SECC (highlighted with red box in Fig. 5). From the protocol point of view, the *client-server* model is used, where the EVCC takes the role of the *client* of the protocol, initiating the communications, and the SECC

takes the role of the *server*. The EVSE-SECC interface is out of the scope of the standardization and is handled in the WAS by the SECC-Session-Manager. The SECC SUT has a dummy SECC-Session-Manager, which does not rely on an actual EVSE. This allows us to focus on the relevant part for conformance tests, without relying on an EVSE and an EV to run the tests while validating and verifying major parts and aspects of the implementation. The conformance test cases are described leveraging this test architecture and are specified in TTCN-3 Core Language for ISO/OSI Network Layer (Layer 3) and above. Note that underlying protocols, such as UDP, TCP/TLS, etc., are not tested directly during the conformance tests; however, the test framework relies on them.

Nowadays TTCN-3 [10] is widely used as a testing language for standards in telecommunications, and it is even used in ITS. TTCN-3 is mostly applied for protocol testing but other test areas (software, system, etc.) and verification objectives (interoperability, robustness, etc.) are starting to use it. In this paper, TTCN-3 is used for the test cases implementation of protocols of ISO/IEC 15118-2-ED2.

B. Test System Implementation

The presented V&V test tool consists of several applications, configuration files, test scripts and run-time environment that run on an embedded computer. Test scenarios and cases are described in TTCN-3 language that is compiled into a binary program, the so-called ETS (executable test suite).

The TTCN-3 scripts are obtained from ISO/IEC 15118-4:2018 [5]. However, some parts are modified since these test scripts are written for wired case, not for wireless (e.g. absence of PLC) and ISO/IEC 15118-9 currently does not exist. These scripts are written in a specific script language (TTCN-3). The V&V test tool contains two main components. These are given as follows: TTCN-3 complier/executor and Test ports (TPs).

These components and their subcomponents are depicted in Fig. 6. The used TTCN-3 compiler and executor is the open source Titan TTCN-3 compiler developed by Ericsson [14]. Titan is a TTCN-3 compilation and execution environment with an Eclipse-based IDE. The Test executor requires a configuration file (cfg). This file contains the input parameters (e.g. the group of test cases that should be executed, use TLS or not etc.). The Titan compiler builds an executable (binary) test suite (ETS) from the TTCN-3 scripts, the test port code and the Titan runtime library. Note that it is not mandatory for ETS to be executable. Titan allows very flexible runtime parameterization of the test cases (e.g. IP addresses, port numbers etc.). The values of runtime parameters need not to be defined at development time, however, default values can be specified, but they can be provided just before the test execution session. In this way, flexible execution scenarios can be created without re-building the ETS.

The TTCN-3 code is generic, therefore the interfaces between the tester and the tested entity (i.e. SUT) are specified at the level of the exchanged abstract data messages and signals. Setting up and maintaining the transport connections and sending/receiving "real" messages and signals are the tasks of interface adaptors. Adaptors are called test ports (TPs) and are plugins written in C/C++ (as illustrated in Fig. 6).

Note that ETS can run on a traditional PC or on a mini/embedded PC, only a Linux environment is required, with the package of OpenSSL (in case of TLS). The computational capacity is usually not a bottleneck for such TTCN-3 based black box testing.

After executing the ETS (with the proper configuration file) the results of test cases are visible for the user by parsing the log file manually or via a graphical interface. The ETS is responsible for assembling the test packets, which are then injected into the network, and transmitted to the SUT that is the SECC in this case. Based on the response from the SUT or even on the existence of the response taking into account time restrictions as well, verdict is made and presented to the test engineer via a suitable, graphical user interface.

The test ports should take care of the following. In the case of SDP, it shall insert a V2GTP header to the SDP message; remove and process the V2GTP header from received message and send/receive UDP multicast message to/from SUT over IPv6.

In the case of V2G application layer message exchange the test port shall insert a V2GTP header to the V2G message; remove and process the V2GTP header from received message; encode/decode message with EXI, encrypt/decrypt V2G message (if TLS is enabled) and send/receive TCP message to/from SUT over IPv6.

C. Test Configuration

The main parameters of the test configuration – used by the test cases – are summarized in Table I.

D. Test Cases and Validation

Black box testing is used in this manuscript to test the SECC implementation in the WAS. This method of testing examines the behaviour of a SUT without considering the internal implementation and structure of the SUT, thus relying on the SUT's open interface for testing. The test tool acts as an EVCC and sends SDP/V2G requests to SECC. The SECC shall respond to them in time.

TABLE I
A COLLECTION OF CONFIGURATION PARAMETERS USED BY TEST CASES

Parameter	Value	Description
LogFile	NeMo-TUB-BIT-%n.log	The filename and path of the log file.
LogSourceInfo	yes	The tool should log the source information also.
PIXIT_SECC_CMN_TLS	false	Use TLS in the V2G communication.
PICS_CMN_CMN_V2gtpSdp	true	Test the SUT with V2GTP-SDP test case set.
PICS_CMN_CMN_Sdp	true	Test the SUT with SDP test case set.
PICS_CMN_CMN_SupportedAppProtocol	true	Test the SUT with V2G SupportedAppProtocol test case set.
PICS_CMN_CMN_SessionSetup	true	Test the SUT with V2G SessionSetup test case set.
pt_V2G_UDP_SDP_Port.debugging	yes	The log should contain UDP test port debug messages.
pt_V2G_UDP_SDP_Port.multicastAddress	"ff02::1"	IPv6/UDP multicast address used by SDP request.
pt_V2G_UDP_SDP_Port.multicastPort	15118	IPv6/UDP multicast port used by SDP request.
pt_V2G_UDP_SDP_Port.ifind	“eth0”	The index of the network interface.

TABLE II
DEMONSTRATION OF THE TEST CASES

Test Case identifier	Test objective	Expected behavior of SUT
TC_SECC_V2GTPSDP_001	The V&V test tool sends a "***SECCDiscoveryReq***" message with the V2GTP header information "protocolVersion" equals '0x01'H, 'invProtocolVersion' equals 'FE'H and 'payloadType' equals '0x8001'H. (V2GTP Header is matched for V2G message content).	Test System then checks that the SUT sends a "***SECCDiscoveryRes***" message with the V2GTP header, information 'protocolVersion' equals '0x01'H, 'invProtocolVersion' equals 'FE'H and 'payloadType' equals to '0x8001'H.
TC_SECC_SDP_001	The V&V test tool sends a "***SECCDiscoveryReq***" message with 'Security' equals '0x10'H and 'TransportProtocol' equals to '0x00'H.	V&V test tool then checks that the SUT sends an "***SECCDiscoveryRes***" message with 'Security' equals '0x10'H, 'TransportProtocol' equals '0x00'H and a valid port and IP address.
TC_SECC_V2G _001	The V&V test tool sends a "***SupportedAppProtocolReq***" message with a list of valid AppProtocols including ISO namespace and all additional mandatory parameters.	The V&V test tool then checks that the SUT sends a "***SupportedAppProtocolRes***" message with response code *'OK_SuccessfulNegotiation'* or *'OK_SuccessfulNegotiationWithMinorDeviation'* and all additional mandatory parameters.

The SDP protocol uses UDP (on a fixed port 15118) for communication, meanwhile V2G uses TCP/TLS dynamic ports between the ranges of 49152 – 65535.

Since the SECC implementation of WAS does not cover all the V2G message set given in [3], and due to page limitations only three test cases are shown here.

In order to demonstrate the capabilities of the V&V test tool we choose three test cases, with each of them belonging to a dedicated protocol (i.e. V2GTP, SDP and V2G application layer message exchange). The objective of the test cases and the expected behaviour is collected in Table II.

1) V2GTP – V2G Transfer protocol

V2GTP is responsible for the encapsulation of an SDP Discovery Request or any V2G application layer message. The V2GTP message consists of two parts, the header and the payload. The payload contains the pure SDP or V2G message, meanwhile the V2GTP header contains information about the protocol version, the payload type and size.

2) SDP – SECC Discovery protocol

The SDP protocol is responsible for the SECC discovery and the negotiation of the transport protocol (i.e. to encrypt the transport layer messages). In this test the V&V tool sends the SDP request and the SUT shall respond with a valid and adequate SDP response. See Table II for further details.

3) V2G – SupportedApplication message exchange

The first V2G application layer message is entitled as "Supported Application Request", which is also a negotiation message between the SECC and the EVCC in order to decide the V2G protocol version and other parameters. This message is an XML message encoded in EXI format. In this test the V&V tool sends the V2G "SupportedAppProtocolReq" request and the SUT shall respond with a valid and adequate response. See Table II for further details.

E. Test evaluation and results

The results of the tests are presented in this subsection. During the tests, requests were sent to the SUT and the corresponding responses were investigated. The time constraints of the V2G protocol is also taken into account by the test tool. The captured message structure of the given protocol (i.e. SDP, V2G) is illustrated by the Titan's Eclipse IDE log viewer and with Wireshark packet sniffer. From Fig. 7, one can see the V2GTP header. The most important part of the header is the field of protocol type with the value of '9000'H. This value denotes that the payload contains a SECC Discovery request message. Note that the payload size is two bytes.

This message is sent as an UDP multicast message to the IPv6 address of "ff02::1" on the port of 15118 by the V&V test tool (as depicted in Fig. 7). The SUT will receive the UDP multicast message and answer it with a dedicated (i.e. non-multicast) message.

Note that the payload part shall contain the applied transfer protocol ID and the security layer ID related to SDP. In this case, a simple TCP connection was used without encryption

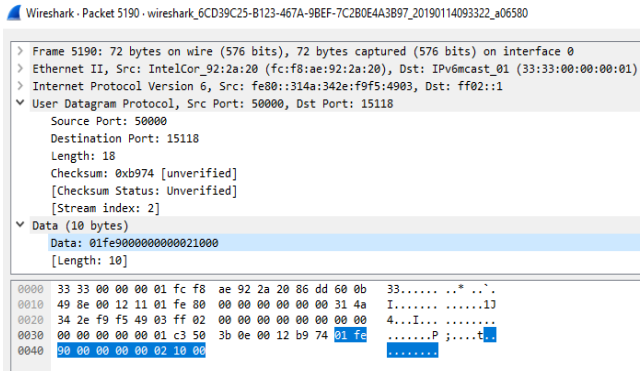


Fig. 7: Sent SECC Discovery Request message

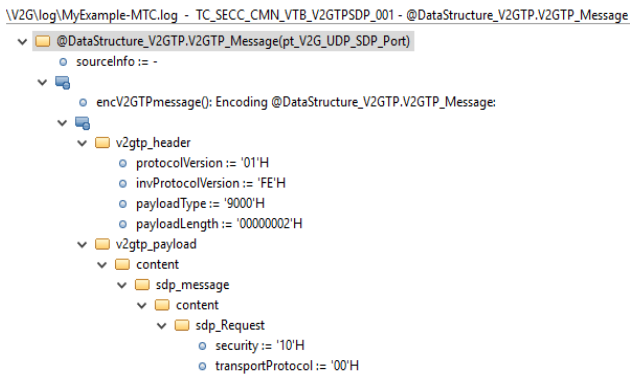


Fig. 8: Structure of the SECC Discovery Request message

(i.e. without TLS layer). Nevertheless, the V&V test tool is capable to establish TLS connection with the SUT as well. However, it is not presented here, instead the plain content is provided in the captured package.

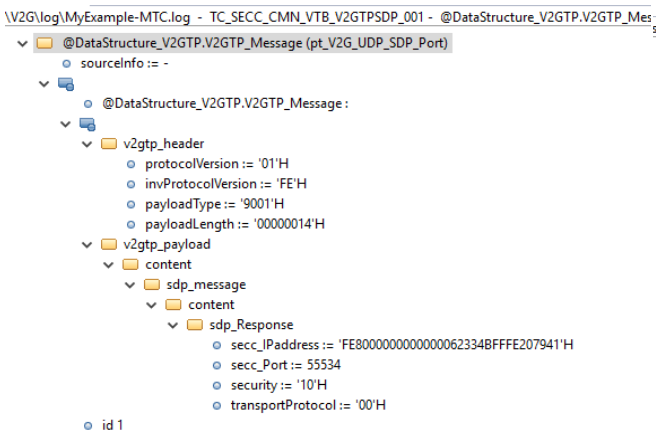


Fig. 9: Structure of the SECC Discovery Response message

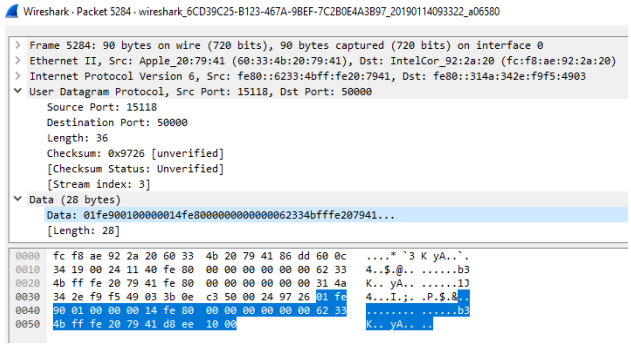


Fig. 10: Received SECC Discovery response message

The desired SDP response (from the SUT) is illustrated in Fig. 8 and in Fig. 9, respectively. The interesting part of the message is related to the V2GTP protocol. In the header, the payload type is '9001'H, which denotes that the payload is a SDP response message (corresponding with the standard [3]). The payload size is 28 bytes, since the SDP response (V2GTP payload part) contains the link-local address of the SUT and the dynamic port, where the next (V2G application layer) message shall be sent. Furthermore, it contains the same values of the field security and transport protocol that was sent in the SDP request. According to Fig. 9, the SDP process completed successfully, thus the V2G application layer message exchange begins. The V&V test tool first sends an EXI encoded message to the SUT containing the parameters depicted on Fig. 11. Afterwards it starts the timer and waits for the SUT's response. This request message contains protocol namespace and the versions supported by the EVCC (in our case the test tool). The SECC (the SUT in this case) should respond that the proposed protocol version is supported by the SECC or not. If it is supported, then the SUT shall send an adequate response, containing a response code. Otherwise, the SUT ends a failed response code, to inform the EVCC that the proposed version is not supported.

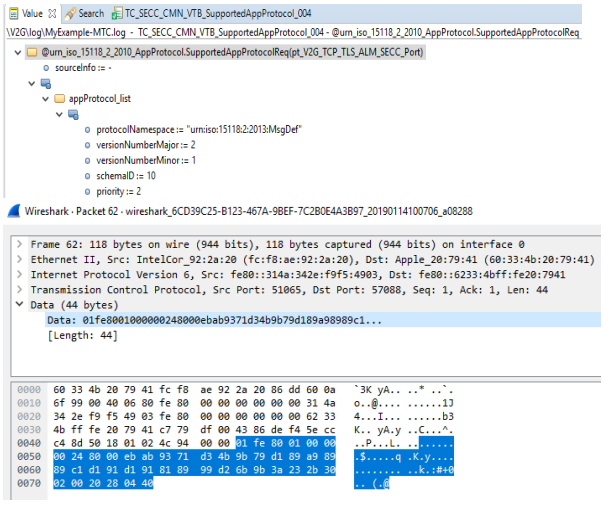


Fig. 11: V2G Supported Application Protocol request message

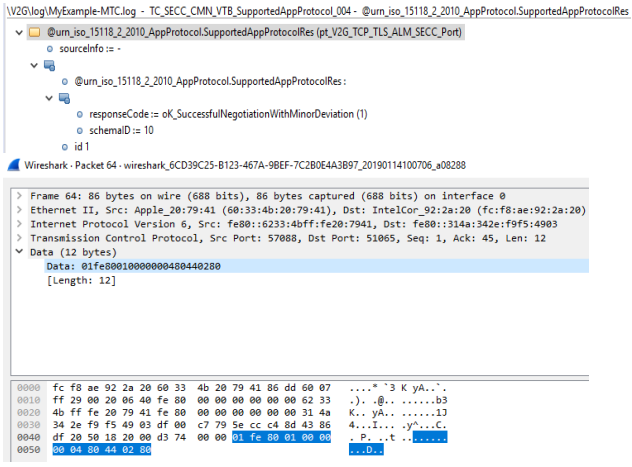


Fig. 12: V2G Supported Application Protocol response message

This message pair is illustrated in Fig. 11 and Fig. 12, respectively. From the incoming response, the V&V test tool first checks the V2GTP header. In this case, the payload type should be '8001'H. Thereafter, from the payload the V&V test tool is able to decode the EXI stream and collect all the necessary information. In this example, the protocol version offered by the EVCC (test tool) is supported by the SUT, thus the negotiation is successful and the response code is *'OK_SuccessfulNegotiationWithMinorDeviation'*.

The results of three test cases are collected in Table III. All the three test cases are evaluated through wireless link and all of them ended with the verdict "Pass".

TABLE III
TEST RESULTS

Test Case identifier	Verdict	Comments
TC_SECC_V2GTPSDP_001	Pass	V2GTP Header message was correct. (SDP Response Message).
TC_SECC_SDP_001	Pass	SDP Response message was correct.
TC_SECC_V2G _001	Pass	SupportedAppProtocolRes message was correct

V. CONCLUSION

In this paper a wireless authentication solution prototype has been presented, which allows electric vehicle owners to identify themselves nearby the charging station, but before connecting the plug to the EVs. Furthermore, we built a conformance test system for the SECC in accordance with the ISO/IEC 15118 standards. The conformance tests are evaluated with a TTCN-3 framework. The main advantage of the proposed V&V test tool is that it is configurable and extendable, therefore subsets of V2G message exchanges can be also executed, or new TTCN-3 test cases can be added next to the conventional ones.

In the manuscript, three test cases were introduced from the developed set for illustration purpose. From the designated tests, it is approved that SDP and V2G communication is possible via wireless links.

The possible future works include the followings. The Test tool covers the V2G messages given in the current version of ISO 15118-2. In the next version of this standard (expected at the end of 2019 or mid 2020) will have more V2G messages, which are related to bidirectional- and wireless charging. One possible extension of the test tool is to support those new message pairs. The current version of the V&V tool uses an Eclipse plugin. In the future, it is desired to have a dedicated graphical user interface (GUI). Another possible extension is to support other charging related protocols, next to V2G, like OCPP (Open Charge Point Protocol).

VI. REFERENCES

[1] M. Mültin, "ISO 15118 as the Enabler of Vehicle-to-Grid Applications," 2018 International Conference of Electrical and Electronic Technologies for Automotive, Milan, 2018, pp. 1-6.

[2] ISO 15118-1:2013: Road vehicles -- Vehicle-to-Grid Communication Interface -- Part 1: General information and use-case definition

[3] ISO 15118-2:2014: Road vehicles -- Vehicle-to-Grid Communication Interface -- Part 2: Network and application protocol requirements

[4] ISO 15118-3:2015: Road vehicles -- Vehicle to grid communication interface -- Part 3: Physical and data link layer requirements

[5] ISO 15118-4:2018: Road vehicles -- Vehicle to grid communication interface -- Part 4: Network and application protocol conformance test

[6] ISO 15118-8:2018: Road vehicles -- Vehicle to grid communication interface -- Part 8: Physical layer and data link layer requirements for wireless communication

[7] ISO 15118-6:2015: Road vehicles -- Vehicle to grid communication interface —Part 6: General information and use-case definition for wireless communication

[8] O. Simon and D. Shkadarevich, "Application of V2G communication for wireless interoperable power transfer," 2017 Twelfth International Conference on Ecological Vehicles and Renewable Energies (EVER), Monte Carlo, 2017, pp. 1-5.

[9] A. Krivchenkov and R. Saltanovs, "Analysis of wireless communications for V2G applications using WPT technology in energy transfer to mobile objects," 2015 56th International Scientific Conference on Power and Electrical Engineering of Riga Technical University (RTUCON), Riga, 2015, pp. 1-4.

[10] ITU-R Z.161 : Testing and Test Control Notation version 3: TTCN-3 core language, https://www.itu.int/rec/T-REC-Z.161/

[11] K. Hänsch et al., "An ISO/IEC 15118 conformance testing system architecture," 2014 IEEE PES General Meeting | Conference & Exposition, National Harbor, MD, 2014, pp. 1-5.

[12] S. Gröning, C. Lewandowski, J. Schmutzler and C. Wietfeld, "Interoperability Testing Based on TTCN-3 for V2G Communication Interfaces," 2012 International Conference on Connected Vehicles and Expo (ICCVE), Beijing, 2012, pp. 298-303.

[13] Minho Shin, Hwimin Kim, Hyoseop Kim 1 and Hyuksoo Jang, "Building an Interoperability Test System for Electric Vehicle Chargers Based on ISO/IEC 15118 and IEC 61850 Standards", Applied Sciences Vol.6, No.6:165

[14] Ericsson's Titan TTCN-3 compiler, https://projects.eclipse.org/projects/tools.titan

[15] El Sayed, N, Lan L L. I. N., Goa, F., & Shi, X."eCo-FEV: efficient Cooperative infrastructure for Fully Electric Vehicle." IEEE Transportation Electrification, April 2014.

[16] N. El Sayed, "A Prototypical Implementation of an ISO 15118 Based Wireless Vehicle to Grid Communication for Authentication over Decoupled Technologies" 2019 International Conference of Electrical and Electronic Technologies for Automotive, Torino, 2019.

[17] OpenEVSE tool kit, https://www.openevse.com/kits.html

Ádám Knapp received his M. Sc. degree in software engineering in 2011 from the Budapest University of Technology and Economics, Hungary. He is a member of IEEE. At the present, he is software developer at BroadBit Hungary Kft. and assistant research fellow at BUTE, Dept. of Networked Systems and Services. His main working area includes communication theory, 4/5G mobile networks and cooperative intelligent transportation systems.

Zoltán Jakó received his M.Sc. degree in electrical engineering, from Budapest University of Technology and Economics (BUTE), Budapest, Hungary, in 2011. He received the Ph.D. degree in the Department of Networked Systems and Services, BUTE at 2017. He is a member of IEEE. He is software developer at BroadBit Hungary Kft. and assistant research fellow at BUTE, Dept. of Networked Systems and Services since 2011. His research interests includes network design with stochastic geometry, next-generation heterogeneous network analysis, vehicle-to-vehicle (V2V) communication and vehicle-to-grid (V2G). He has been the involved several EU FP7 and Horizon2020 research projects.

Nadim El Sayed graduated with distinction from the Technische Universität Berlin (TUB) in Computer Engineering with specialization on Telecommunication Networks in 2010, and received a distinction award from the VDI in 2011. Since then he has been a researcher at the DAI-Labor, TUB in the Competence Center for Networks and Mobility. He was involved in many national and EU projects working in the areas of Mobile Communication, E-Mobility, Smart Grids, IoT and Industry 4.0. He has published scientific journal papers, conference articles and book chapters, and authored open source software for smart-metering and power quality monitoring. In 2018 he was an invited expert speaker at the German federal Ministry of economy regarding the directive 2014/94/EU2014 on the deployment of alternative fuels infrastructure.

An OFDMA-based Hybrid MAC Protocol for IEEE 802.11ax

Gazi Zahirul Islam and Mohammod Abul Kashem

***Abstract*—Two types of MAC mechanisms i.e., random access and reservation could be adopted for OFDMA-based wireless LANs. Reservation-based MAC is more appropriate than random access MAC for connection-oriented applications as connection-oriented applications provide strict requirements of traffic demands. On the other hand, random access mechanism is a preferred choice for bursty traffic i.e., data packets which have no fixed pattern and rate. As OFDMA-based wireless networks promise to support heterogeneous applications, researchers assume that applications with and without traffic specifications will coexist. Eventually, OFDMA-based wireless LAN will deploy hybrid MAC mechanisms inheriting traits from random access and reservation. In this article, we design a new MAC protocol which employs one kind of hybrid mechanism that will provide high throughput of data as well as maintains improved fair access policy to the medium among the terminals. The protocol works in two steps, where at step 1 sub-channels are approximately evenly distributed to the terminals and at step 2 terminals within in a sub-channel will contend for medium randomly if the total number of terminals of the system is larger than the number of sub-channels. The details of the protocol is illustrated in the paper and we analyze the performance of our OFDMA-based multi-channel hybrid protocol using comprehensive computer simulations. Simulation results validate that our proposed protocol is more robust than the conventional CSMA/CA protocol in terms of throughput, collision reduction and fair access. In addition, the theoretical analysis of the saturation throughput of the protocol is also evaluated using an existing comprehensive model.**

***Index Terms*—Throughput, MAC, OFDMA, IEEE 802.11ax, CSMA/CA, Wi-Fi 6.**

I. INTRODUCTION

The rapid growth of demand for high-speed WLAN has driven the exhaustive research to enhance the throughput by employing a variety of medium access control (MAC) mechanisms. The efficiency of MAC plays a major role to enhance the throughput of any wireless LAN system. One of the innovative and promising access procedures for MAC is orthogonal frequency division multiple access (OFDMA) which originally derived from orthogonal frequency division multiplexing (OFDM). An OFDMA system uses a group of non-overlapping sub-carriers to form a sub-channel that can be allocated to each transmitting station. Thus, multiple stations can send data concurrently without having collision [1]. Absorbing the advantages of OFDM, OFDMA-based MAC protocol can further enhance efficiency by increasing multiuser diversity. As such superiority of OFDMA technology, some wireless systems such as WiMAX leverages it from the very beginning.

According to the functional requirements of IEEE 802.11ax, Wi-Fi should achieve at least 4 times improvement in the average throughput per station (terminal) as well as should support highly dense systems [14]. The physical data rate in Wireless LAN has been remarkably boosted due to more available bandwidth resources and the arrival of modern technologies such as MIMO [5]. However, the MAC layer of Wireless LANs has not changed significantly for the last 16 years. Since their birth, Wireless LANs employed distributed coordination function (DCF) as the MAC layer protocol [19]. According to the DCF protocol, only one station can utilize the channel resource and send data at the same time [9]. DCF rules employed in IEEE 802.11 are suited to sparsely dense Wireless LAN environment, while in the highly dense system the MAC efficiency of DCF would be very poor due to the provision of single user accessibility [15]. To overcome the difficulties mentioned above, multiuser MAC is required instead of a single user [16]. Since Wireless LANs have already included OFDM as modulation technology, OFDMA technology is highly recommended for next generation Wireless LANs [17]. An OFDM adopted system enables a single terminal to utilize all the sub-channels at any given time while an OFDMA adopted system enables multiple terminals to use a different set of sub-channels, thereby providing concurrent transmission of more than one terminal [2].

Two types of MAC mechanisms namely, random access and reservation can be employed for OFDMA-based wireless LANs. Reservation-based MAC is more appropriate than random access MAC for connection-oriented applications as connection-oriented applications provide clear specifications of traffic demands. Reservation-based MAC ensures graceful support for Quality of Service (QoS). However, it is not appropriate for applications that contain no traffic specifications. For example, in data networks like the Internet, an application is usually characterized by bursty traffic, i.e., data packets arrive in an arbitrary pattern and rate. So, it would be unwise to reserve a certain amount of resources (e.g., sub-channels in OFDMA) for applications in a data network. Thus, random access mechanism is a preferred choice for bursty traffic. As OFDMA-based wireless networks promise to support heterogeneous applications, it is anticipated that applications with clear traffic specifications and those without traffic specifications will coexist [3]. To this end, both

Gazi Zahirul Islam is currently pursuing his PhD at Bangladesh University of Professionals, Dhaka-1216, Bangladesh. He is also teaching at Department of Computer Science and Engineering, Daffodil International University, Bangladesh (e-mail: zahircuet@gmail.com).

Mohammod Abul Kashem is a Professor of Department of Computer Science and Engineering, Dhaka University of Engineering and Technology, Gazipur-1700, Bangladesh (email: drkashemll@duet.ac.bd)

reservation-based and random-access MAC procedures are optimized for an OFDMA-based wireless LAN. In other words, the MAC protocol of an OFDMA-based wireless network will provide a hybrid MAC mechanism for both random access and reservation.

In this paper, we propose an innovative MAC protocol which employs a hybrid mechanism that could provide high throughput of data as well as able to maintain improved fair access policy to the medium among the terminals. The protocol works in two steps, where at step 1 sub-channels are approximately evenly distributed to the terminals and at step 2 terminals within in a sub-channel will contend for medium randomly if the total number of terminals of the system is larger than the number of sub-channels. The details of the protocol will be described in the 'Protocol Illustration' section i.e. Section IV.

The rest of the article is organized as follows. At first, we discuss related works and motivation in Section II. Section III contains the system model and Section IV contains protocol illustration. Mathematical analysis of the saturation throughput of the protocol is evaluated in Section V using a comprehensive model. Simulation is conducted by renowned 'NS-3 Simulator' [10] and presented the result in Section VI. Finally, Section VII concludes the paper.

II. Related Works And Motivation

Recently, there has been rigorous research devoted to the combination of OFDMA with MAC. Xuelin et al. in [6] designed a multi-step slot reservation hybrid MAC protocol named 'TR-MAC' for ad hoc networks which incorporates the strengths of TDMA (Time Division Multiple Access) and DCF of IEEE 802.11. TR-MAC eliminates the slot assignment algorithm, reduces the control packets negotiation and avoids extra contentions. Thus, enhanced the throughput of MAC without incurring additional overhead. The researchers in [7] devised a model named 'CCRM' which innovates a new asynchronous MAC protocol with cooperative channel reservation. Compared with legacy channel reservation MAC protocols, where channel reservation information (CRI) cannot be obtained reliably due to either transmission errors or packet collisions, CCRM improves the reliability of channel reservation using cooperative channel reservation mechanism.

Several random access protocols have been designed for OFDMA-based WLANs. The authors of [8] and [3] proposed a protocol using a two-dimensional backoff scheme to enable the terminals accessing the channel both in the time and frequency domains. The articles [20] and [21] divide stations into multiple groups and the stations in the same group share the same sub-channel for channel access. Once the access point receives an RTS (request-to-send) frame from the sub-channels, it replies with a CTS (clear-to-send) frame to assign the channel resources.

Choi et al. [8] put forward an innovative fast retrial slotted ALOHA-based scheme to reduce access delay, but the throughput of the protocol is very poor due to high collision probability. In article [2], the researchers designed a random access model based on the CSMA/CA technique that outperforms traditional ALOHA protocol. According to that model, a terminal employs only one backoff timer for all the sub-channels, and the timer could not reflect different traffic loads in different sub-channels. As a result, the channel utilization efficiency of the model in [2] is still not satisfactory. This constraint is then resolved by Wang in [3], where a terminal employs one backoff timer for each of the sub-channel. Therefore, the transmission status of one sub-channel does not affect the rest of the sub-channels. To overcome the half-duplex limitation of the wireless radio the authors of [3] proposed utilizing an additional radio to sense the medium on all other sub-channels while the original radio is busy in transmission on a certain sub-channel. Thus, the scheme improved transmission concurrency on multiple sub-channels. However, this sort of scheme having a dedicated sensing module is not applicable to the station with a single radio. Jia Xu et al. [4] introduce intermittent carrier sense technique that permits a single-radio OFDMA station to access multiple sub-channels simultaneously. However, the total throughput of the system and the max-min fairness is not yet satisfactory to meet the demand of IEEE 802.11ax network.

Considering above ideas and facts, we design a new MAC protocol named 'HTFA' for high throughput and fair access. The main contribution of HTFA is as follow:

- One of the major goals of IEEE 802.11ax (Wi-Fi 6) is to improve the total system throughput as well as per terminal throughput. HTFA provides higher throughput than several promising protocols which will be described in the 'Performance Evaluation and Simulation' section.
- HTFA leverages hybrid mechanisms to distribute channel access time more evenly among the terminals. Thus it ensures fair access policy and performs better than SRMC-CSMA/CA and CM-CSMA/CA introduced in [4] and [3] respectively.
- The terminals in HTFA will not contend for sub-channel access if the number of terminals is smaller or equal to the number of sub-channels. Thus, the probability of frame collision is zero. Since there is no backoff slot, there is no idle slot as well. Hence, system throughput increases significantly.
- We perform an extensive simulation with network simulator NS-3 [10] which is presented in Section VI. Simulation results confirm validation of our protocol in terms of throughput, collision reduction and fairness.

III. System Model

We consider an OFDMA-employed WLAN where total bandwidth B is equally distributed to M sub-channels. Hence the bandwidth of a sub-channel would be B/M. There are N stations and only one access point (AP) in our system. By choosing different sub-channels, more than one station can communicate with the AP at the same time without suffering from co-channel interference. In such a network collision would occur if and only if multiple stations send packets on the same sub-channel concurrently.

The 802.11ax standard hires some technological developments from 4G cellular technology to support more stations in the same channel bandwidth leveraging OFDMA. 802.11ax not only adopted OFDM digital modulation scheme but also allocates a group of non-overlapping subcarriers to individual stations. The standard partitions the existing 802.11 channels which may be 20/40/80/160 MHz wide into smaller

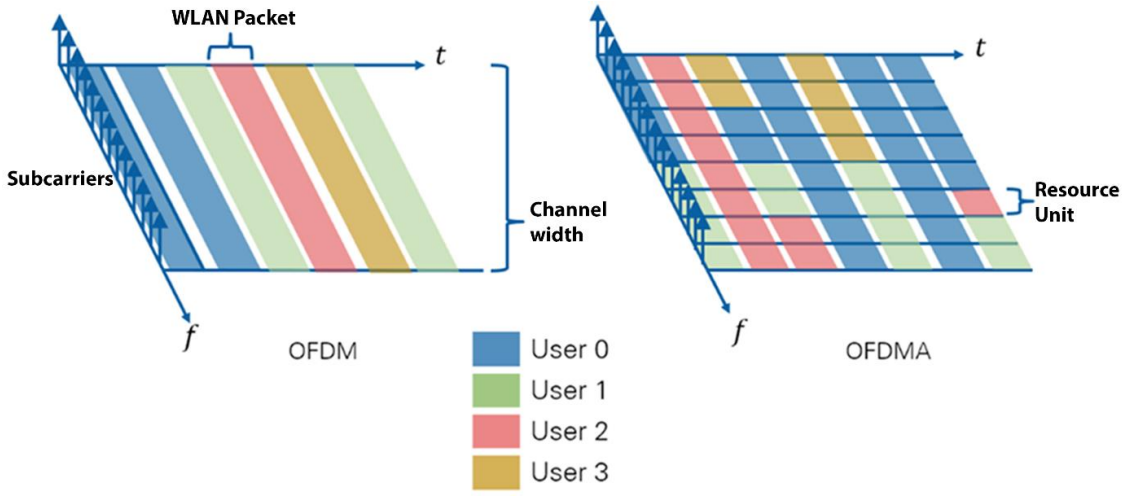


Fig. 1. The contrast between OFDM and OFDMA

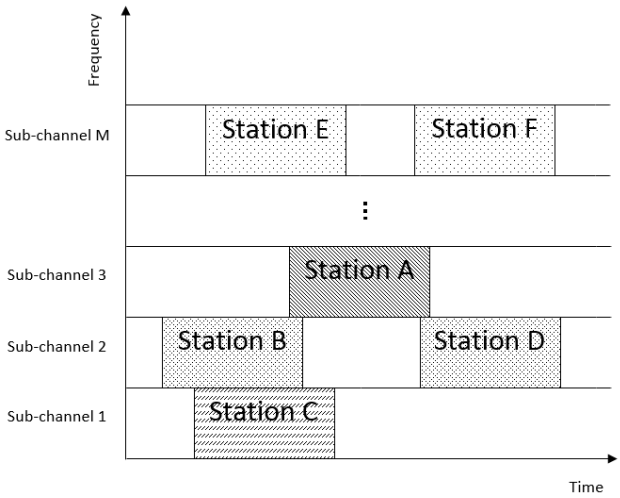


Fig. 2. A 2-dimensional time-frequency access model

sub-channels with a specified number of orthogonal sub-carriers [18]. Following LTE (Long Term Evolution) nomenclature, the 802.11ax standard terms the smallest sub-channel as a resource unit (RU) which contains at least 26 subcarriers.

Observing different user's traffic needs, the access point decides on allocating the channel, always allocating available resource units on the downlink path. The AP might assign the whole channel to only one terminal or it may divide the channel to form sub-channels for serving multiple terminals simultaneously (Fig. 1). In congested areas where lots of terminals would normally compete inefficiently for channel access, the OFDMA technology can now serve them concurrently with a smaller but dedicated sub-channel. As a result, the average throughput per user is enhancing significantly.

As mentioned earlier, in comparison with the single-channel CSMA/CA, the multi-channel system facilitates stations to access the sub-channels simultaneously without having interference. In a multi-channel system, stations can compete for available resources from both time and frequency domain as shown in Fig. 2 [22]. In the time domain, stations could acquire the time slots of a sub-channel when the sub-channel is not busy and avoid possible collision (if more than one station present) using binary exponential backoff (BEB) algorithm. In the frequency domain, stations could use different sub-channels concurrently and prevent interference using OFDMA mechanism.

We suppose every station maintains its own timer and synchronizes its timer with other station's timer. To ensure clock synchronization, the AP informs the reference time information to participating stations at a regular interval according to the time synchronization function suggested by the IEEE 802.11 standard [9]. It is noted that imperfect synchronization generates clock offset among different stations. In the OFDMA system orthogonality among sub-channels cannot be guaranteed if the clock offset is exceeded the threshold [2]. Therefore, we assume synchronization would be maintained efficiently to confine the maximum clock offset within the threshold, thereby ensuring the orthogonality among the OFDMA sub-channels.

IV. PROTOCOL ILLUSTRATION

In this section, at first, we will discuss the method of sub-channel distribution which is a complex procedure comparing to the pure random access mechanism. Then we will discuss the basic access mechanism and at last, we will discuss the advantages of our proposed MAC protocol.

A. Sub-channel Distribution

The main distinguishing feature of our hybrid protocol is its uniqueness in distributing the sub-channels to the terminals. As stated earlier, HTFA works in two steps: step 1 and step 2 where step 2 is conditional. In step 1 sub-channels are approximately

evenly distributed to the terminals and in step 2 terminals within in a sub-channel will contend for medium randomly if the total number of terminals of the system is larger than the number of sub-channels. We said "approximately evenly distributed" because the number of terminals in the sub-channels is differed by at most one. We consider five distinguish scenarios for distributing the sub-channels:

Scenario 1: Number of sub-channels is equal to the number of terminals ($M = N$)

Scenario 2: Number of sub-channels is greater than the number of terminals ($M > N$)

Scenario 3: Some terminals leave the network early

Scenario 4: Number of terminals is greater than the number of sub-channels ($N > M$)

Scenario 5: Some terminals join the network after some time

We describe each scenario as follow:

Scenario 1: Number of sub-channels is equal to the number of terminals ($M = N$)*:* Suppose we have three terminals namely, Station A, Station B and Station C; and three sub-channels namely, Sub-channel 1, Sub-channel 2 and Sub-channel 3. Since in this case $M = N$ and the sub-channels are evenly distributed, each terminal will get exactly one sub-channel (Fig. 3 (a)). It is not possible one terminal gets two sub-channels and other two get one and zero terminal respectively. In this way, HTFA ensures fair access to the medium among the stations which is absent in the article [3] and [4].

Scenario 2: Number of sub-channels is greater than the number of terminals ($M > N$)*:* Suppose we have two terminals Station A and Station B; and three sub-channels as mentioned above. Since in this case $M > N$ and sub-channels are approximately evenly distributed, one terminal gets two sub-channels and other terminal gets one sub-channel i.e. Station A gets one sub-channel and Station B gets two sub-channels (Fig. 3 (b)).

Scenario 3: Some terminals leave the network early: Again, suppose at the beginning, Sub-channel 1, Sub-channel 2 and Sub-channel 3 are assigned to Station B, Station A and Station C respectively. After some time, Station B sent all of its data and releases its sub-channel i.e. Sub-channel 1. Now, one of the two remaining terminals (A or C) can acquire B's sub-channel and send data (Fig. 3 (c)).

Scenario 4: Number of terminals is greater than the number of sub-channels ($N > M$)*:* Now suppose we have four terminals Station A, Station B, Station C and Station D; and three sub-channels as mentioned above. In this case, the first 3 terminals (A, B and C) are assigned to three sub-channels and the fourth terminal (Station D) is assigned to anyone sub-channel. That means, two sub-channels get 1 terminal each and one sub-channel gets 2 terminals. Suppose, sub-channel 2 gets two terminals (Station A and Station D) which is shown in Fig. 4. Now that Station A and Station D are assigned to the same sub-channel (Sub-channel 2), they will randomly contend for channel access according to the legacy DCF (Distributed Coordination Function) rule.

Fig. 4 delineates random access procedure, where initially Station A and Station D generate random backoff number 5 and 7 respectively. As Station A generates a smaller number than Station D, Station A will access the sub-channel first when it's backoff counter reaches to zero. After one more slot time, D's backoff counter reaches to zero and thereby access the sub-channel. In the second round, Station A and Station D generate new backoff value 5 and 2 respectively and the procedure will continue according to the DCF rule.

Scenario 5: Some terminals join the network after some time: How does a terminal join in the network after some time depends on the current status of the network. There are two statuses:

(i) $M > N$
(ii) $N \geq M$

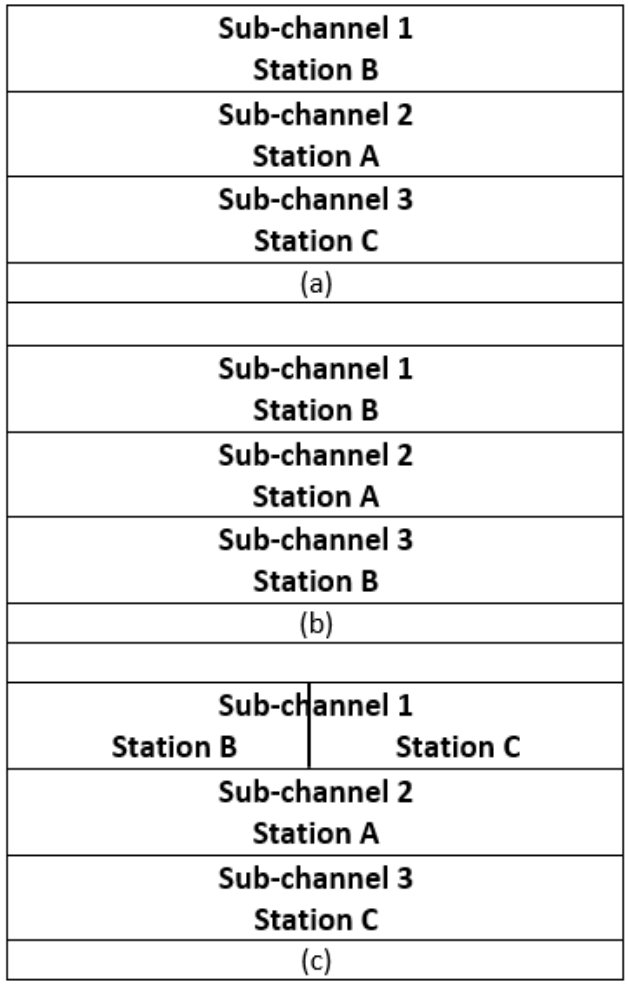


Fig. 3. Sub-channel distribution (a) Case 1 (b) Case 2 (c) Case 3

Fig. 4. The contention of two stations within a sub-channel

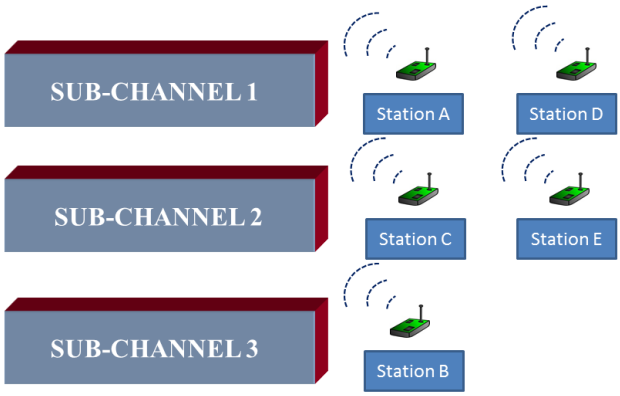


Fig. 5. Network configuration for 3 sub-channels and 5 stations

When M > N, then the new terminal will acquire a sub-channel from the old terminal that has the highest number of sub-channels. The new terminal may need to wait for some time so that old terminal can send its ongoing packet. When N ≥ M, then the new terminal will join a sub-channel that has the minimum number of terminals.

The sub-channel distribution procedure mentioned above is very intuitive. For further clarification of the procedure, we are going to demonstrate an example scenario. Suppose at the beginning there is an access point with no terminals at all and there are three sub-channels SUB-CHANNEL 1, SUB-CHANNEL 2 AND SUB-CHANNEL 3. After some time, Station A joins the network and it will acquire all three sub-channels. Then after some time Station B joins the network and it will seize one sub-channel (e.g. SUB-CHANNEL 3) from Station A. Now Station A has two sub-channels SUB-CHANNEL 1 and SUB-CHANNEL 2; and Station B has one sub-channel (SUB-CHANNEL 3). Again, after some time Station C joins the network and it will seize one sub-channel (e.g. SUB-CHANNEL 2) from Station A. Now each of the three terminals has exactly one sub-channel. Suppose after some time Station D wants to join the network and the AP can put it to any sub-channel (e.g. SUB-CHANNEL 1) to contend for channel access by generating random back-off value. Similarly, Station E arrives and the AP assigns it to the sub-channel which has the minimum number of terminals. Now SUB-CHANNEL 2 and SUB-CHANNEL 3 has a minimum number of terminals (i.e., 1 terminal each) and suppose Station E got SUB-CHANNEL 2. Until now we get network configuration shown in Fig. 5, where SUB-CHANNEL 1 contains Station A and Station D; SUB-CHANNEL 2 contains Station C and Station E; and SUB-CHANNEL 3 contains Station B.

Now suppose Station E wants to leave the network. The sub-channel distribution will not change after leaving Station E since the number of terminals is approximately evenly distributed. After a while, Station C also leave the network. Now the number of terminals is not approximately evenly distributed since SUB-CHANNEL 1 contains two terminals (Station A and Station D) and SUB-CHANNEL 2 contains no terminal. In this case, one terminal of SUB-CHANNEL 1 will migrate to SUB-CHANNEL 2. Thus, each sub-channel gets exactly one terminal and terminals are evenly distributed to the sub-channel.

B. Access Mechanism

A terminal with a new packet to transmit must be associated with the access point (AP). AP will assign sub-channels to the terminals according to the five scenarios mentioned in Section IV-A i.e. sub-channel distribution. Access method depends on the number of terminals available in a sub-channel. There would be two cases:

i. Multiple terminals in a sub-channel
ii. Single terminal in a sub-channel

i. Multiple terminals in a sub-channel: In this case, four-way handshaking (RTS/CTS, DATA/ACK) will be used as shown in Fig. 6 [22]. After DIFS (distributed inter-frame space) interval sending terminal enters into backoff interval. When backoff value reaches to zero, sending terminal transmits RTS (request-to-send) frame to the receiving terminal. After SIFS (short inter-frame space) interval receiver responds with the CTS (clear-to-send) frame. Upon receiving the CTS frame, the sender waits for the SIFS interval and then send the DATA frame. After receiving the DATA frame, the receiver again waits for SIFS interval and then responds with ACK (acknowledgement) frame.

In this case, all stations have to generate a random backoff time before sending data in order to minimize the probability of collision with the frames being sent by other stations in the same sub-channel. If two or more stations generate the same backoff number, a collision would occur and will lose data of the colliding stations as shown in Fig. 7.

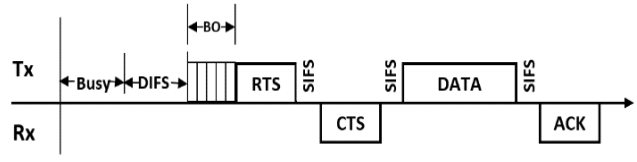


Fig. 6. Four-way handshaking for multiple stations in a sub-channel

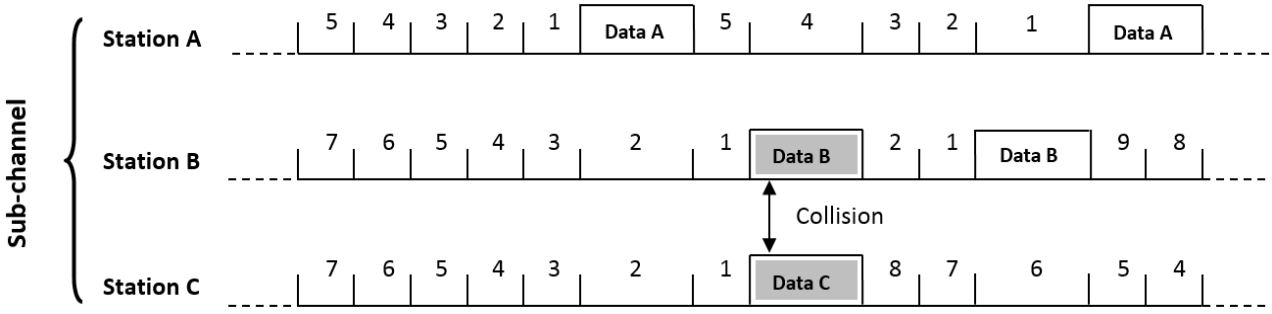


Fig. 7. A Collision between Station B and Station C in one particular sub-channel

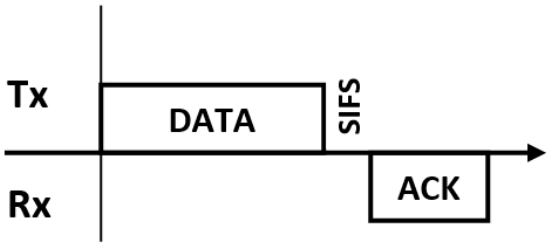


Fig. 8. Two-way handshaking for a single station in a sub-channel

Binary exponential backoff (BEB) algorithm determines the backoff time which is uniformly chosen in the range of [0, W−1] for a contention window size W. At the first transmission attempt W is set to the minimum value W_{min} and is doubled at each backoff stage up to the maximum value $W_{max} = 2^{\alpha}W_{min}$ after each unsuccessful transmission, where α denotes the number of backoff stages.

ii. Single terminal in a sub-channel: In this case, two-way handshaking (DATA/ACK) will be used as shown in Fig. 8. Since only one terminal monopolizes the sub-channel, there is no room for the hidden terminals to participate in the channel sharing. As a consequence, we need not the RTS/CTS frame pair. There is no DIFS and backoff interval as well which reduces overheads and increases throughput. In this case, the sending terminal first transmits the DATA to the receiving terminal. After waiting for a SIFS interval, the receiver sends the ACK frame to the sender.

C. Advantages of HTFA

Throughput Enhancement: The IEEE 802.11ax [also known as High Efficiency WLAN (HEW)] has a challenging goal of increasing the average throughput per user four times in highly dense environment. HTFA provides more throughput than several promising protocols which will be shown in Section VI. The terminals in HTFA will not contend for sub-channel access rather than sub-channels are dedicated to the terminals if $N \leq M$. In this case, terminals monopolize the sub-channel access and there is no random backoff slot which significantly increases the throughput.

Collision Reduction: According to our model, a sub-channel could get at most one terminal when $N \leq M$. Thus, the probability of frame collision would be zero. Hence, the average throughput per terminal as well as total system throughput increases significantly. There is a probability of collision in some sub-channels if and only if $N > M$. However, in any case, the probability of collision in HTFA is smaller than any non-OFDMA channel (i.e. single channel).

Fair Access: Our proposed protocol employs a hybrid mechanism to distribute the sub-channels among the terminals. Thus, it not only provides a high throughput of data but also maintains improved fair access policy to the medium. The protocol works in two steps, where at the first step sub-channels are approximately evenly distributed to the terminals. Thus, the number of terminals in the sub-channels is differing by at most one. In the second step, terminals within a sub-channel will contend for medium randomly when $N > M$. Hence, HTFA performs better than SRMC-CSMA/CA and CM-CSMA/CA introduced in [4] and [3] respectively in terms of both throughput and fairness.

V. THOROUGHPUT ANALYSIS

To measure the efficiency of a protocol, it is expected to validate the protocol employing an appropriate mathematical model that eventually increase the credibility and acceptability of the conducted research. Here, we will investigate the throughput of our proposed protocol using the analytical model presented in [12] by G. Bianchi. Bianchi formulates an ideal model for analyzing the saturated throughput of DCF which is followed by some other researchers. He designed the model using a discrete-time Markov chain, where the backoff mechanism is regulated by conventional single-channel CSMA/CA. Several other papers including [13] which are considered as extensions of [12] investigated the enhanced mathematical model for the actual backoff mechanisms by considering the existence of anomalous slots. Pioneer model in [12] and some of its extensions estimates saturation throughput of the single-channel terminal employing CSMA/CA mechanism. However, we will evaluate the saturation throughput of multi-channel terminals rather than the single channel which are also regulated by the CSMA/CA protocol.

Again, we assume that we have only one Basic Service Set (BSS) and the access point (AP) located at the centre of BSS. The BSS has N terminals and M sub-channels. We also assume that the number of terminals is very much larger than the number of subchannels i.e. $N \gg M$. The concept of saturated condition is the same as in [12] when WLAN carries the maximum load. We suppose that every station always has some packets available for sending. It implies the input queue of each terminal in WLAN is always non-empty in the saturated stage.

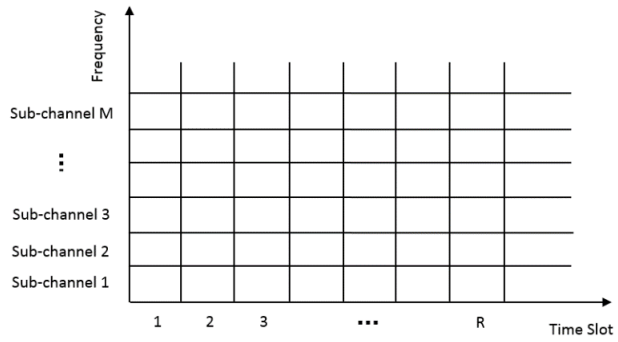


Fig. 9. Time-frequency block

It is not feasible to expand the Markov chain model of multi-channel CSMA/CA from the traditional single channel-based model. Equations in [12] show that the author analyzed the characteristics of a single terminal in WLAN and received the transmission probability (τ) of a packet in a randomly chosen slot time. As a consequence, instead of directly evaluating the multichannel protocol model we will evaluate the probability of successful transmissions from the side of subchannels based on the random backoff mechanism.

In Fig 9 we defined a time-frequency block where each time slot in every sub-channel is considered as a resource utilized for acquiring the medium and transmission of data. We can find the number of time slots *(R)* for the successful transmission of average packet payload (*E[p]*) as $R = T_{total}/T_{slot}$, where T_{total} denotes the total average time and T_{slot} denotes a single slot time. When a terminal gains a particular sub-channel at a random time slot then successful transmission interprets that only this particular terminal acquires this particular sub-channel for *R* time slots. We can get the probability of successful access in one sub-channel at each time slice as $p_k = 1/R$ due to the equal probability of each single time slice under the presumption of the saturated stage.

We define the successful transmission probability P_{jsuc} at the j^{th} sub-channel when a terminal successfully utilizes one sub-channel at one time slot as:

$$P_{jsuc} = \sum_{k=0}^{R-1} p_k \binom{N}{1} \tau(1-\tau)^{N-1}. \quad (1)$$

In the saturation stage, probability of collision P_{jcol} in one sub-channel must be:

$$P_{jcol} = 1 - P_{jsuc}. \quad (2)$$

Since each terminal can only gain one sub-channel at any time slot to send data at the saturated stage, the probability is dependent between P_{1suc} and P_{2suc}. We can find the probability of *1st* and *2nd* sub-channels having successful transmissions as below:

$$P(1suc, 2suc) = P_{1suc} P_{2suc}|_{1suc} \quad (3)$$
$$= P_{1suc} \sum_{k=0}^{R-1} p_k \binom{N-1}{1} \tau(1-\tau)^{N-2}.$$

We define $P_s(i)$ and $P_c(i)$ as there are i sub-channels in our system having successful transmissions and collisions respectively.

Obviously $P_s(1) = P_{1suc}$. In general, the probability of i sub-channels having successful transmission computes as follow,

$$P_s(i) = \binom{M}{i} P(1suc, 2suc, \dots, isuc)$$
$$P\big((i+1)col, (i+2)col, \dots, M\,col\big), \quad (4)$$

We can get $P\big((i+1)col, (i+2)col, \dots, M\,col\big)$ following the procedure of equation (3).

As we assume $N \gg M$, P_{1suc} and P_{2suc} are independent and also equal to each other. Thus, $P(1suc, 2suc)$ could be rewritten as:

$$P(1suc, 2suc) = P_{1suc}\, P_{2suc} = P_{1suc}^2.$$

Simplifying equation (4) yields,

$$P_s(i) = \binom{M}{i} P(1suc, 2suc, \dots, isuc)$$
$$P\big((i+1)col, (i+2)col, \dots, M\,col\big), \quad (5)$$
$$= \binom{M}{i} P_{1suc}^i (1 - P_{1suc})^{M-i}, \quad (6)$$

We can use $P_s(i)$ to find the average number of sub-channels (E_s) having successful transmissions at any single time slot as below:

$$E_s = \sum_{i=0}^{M-1} i\, P_s(i). \quad (7)$$

Finally, we can obtain our desired saturation throughput (S) as follow:

$$S = \frac{E[p].E_s}{T_{slot}}. \quad (8)$$

VI. PERFORMANCE EVALUATION AND SIMULATION

In this section, we first analyze the HTFA protocol with respect to different simulation parameters. After that, we compare and contrast the efficiency of HTFA (High Throughput and Fair Access) with some other promising protocols proposed by different researchers. All the simulation scenarios are implemented in 'Network Simulator-3'.

TABLE I
SIMULATION PARAMETERS FOR HTFA EVALUATION

Parameters	Value
Backoff slot duration	50 µs
DIFS duration	110 µs
Packet transmission time	2.5 ms
Minimum contention window size, W_{min}	32 slot
Number of backoff stages, α	6
Number of stations, N	10

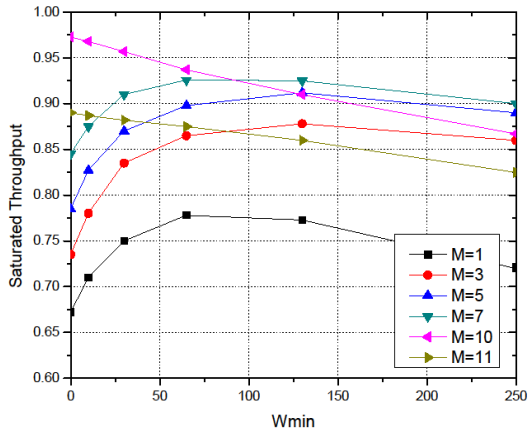


Fig. 10. Saturated throughput with respect to W_{min} for the varying number of sub-channels

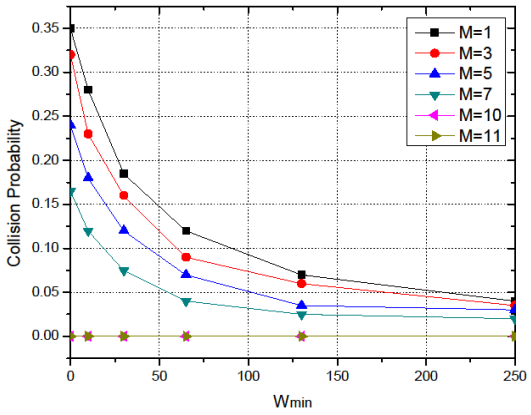


Fig. 11. Collision probability with respect to W_{min} for the varying number of sub-channels

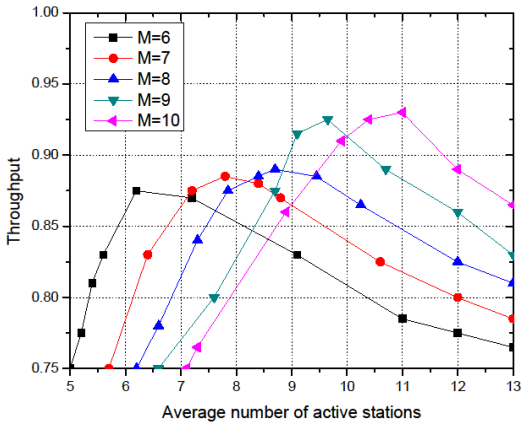


Fig. 12. Throughput in accordance with the average number of participating terminals under non-saturated traffic loads

A. HTFA Evaluation

We analyze the performance of our proposed OFDMA-based multi-channel hybrid protocol through comprehensive computer simulations. The simulation parameters are listed in Table I. In these experiments, we do not consider any transmission failure resulted from channel error.

Fig. 10 shows the saturated throughput of the HTFA protocol with respect to the minimum contention window size (W_{min}) for the varying number of sub-channels. In this experiment we use 10 stations for different sub-channels i.e. M = 1, 3, 5, 7, 10 and 11. The figure shows that maximum saturation throughput is enhancing gradually until the number of sub-channels increases to the number of stations. The saturation throughput decreases when the number of sub-channels exceeds the number of stations. This happens because the extra sub-channels are not utilized efficiently and there is also channelization overhead. We also observe irregular behaviour for M = 10 and 11. Because in these cases, incrementing in the contention window size also increases the idle time while there is no significant reduction in collision probability. This incident suggests that when the number of sub-channels approaches the number of stations, we need not use a large time-domain backoff for collision resolution purpose as the collision probability would be very low in such stage.

Fig. 11 describes the collision probability of our multi-channel protocol with respect to the minimum contention window size (W_{min}) for varying numbers of sub-channels. The graphs show that collision probability decreases as W_{min} increases. It also reveals that collision probability is decreasing as the number of sub-channels is increasing for particular W_{min}. It happens because increasing the number of sub-channels means less contention for channel access. It is observed that for M = 10 and 11, the collision probability is zero because the number of stations (i.e. N = 10) is less than or equal to the number of sub-channels.

Now we evaluate the throughput performance of the proposed 'HTFA' hybrid protocol under non-saturated traffic loads. Fig. 12 shows the resulting throughput with respect to the average number of active stations. We find the similarly as reveals in saturated traffic load: the throughput enlarges until the number of active stations N exceeds the number of sub-channels M but reduces gradually beyond that because collision occurs more frequently among the participating stations. As throughput loss resulted from extra sub-channels is larger than that caused by the collision, it is expected to keep the number of sub-channels slightly smaller than the average number of participating stations.

According to the above findings, we might adopt a system where the number of sub-channels would be fixed adaptively according to the number of participating stations. Specifically, we might incorporate an adaptive control mechanism in such a way that the access point (AP) first estimates the number of participating stations in the wireless LAN, then determines the number of optimum sub-channels from the estimation, and finally announces the result to the terminals through the control channel using beacon messages.

We further investigate the impact of increasing the backoff slot duration on the performance of the OFDMA-adopted multi-channel hybrid system. We examine the saturation throughput of the OFDMA-employed wireless LAN with respect to the

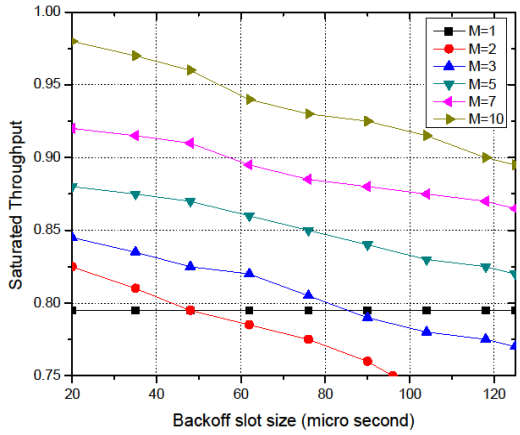


Fig. 13. Saturated throughput with respect to the backoff slot duration

TABLE II
SIMULATION PARAMETERS FOR PROTOCOL COMPARISON

Parameters	Value
Slot time	10 μs
Packet length	1500 bytes
Total channel bandwidth	54 Mbit/sec
Number of stations	3
Number of sub-channels	3
CW_{min}	32
CW_{max}	1024

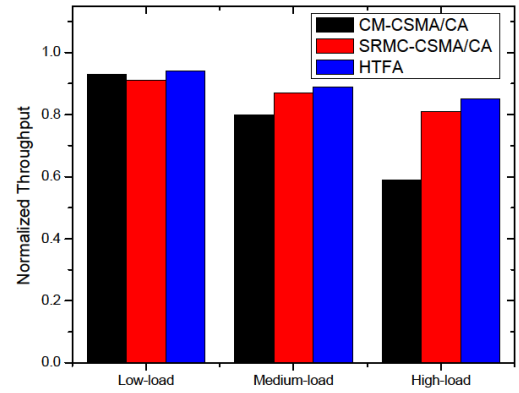


Fig. 14. Normalized throughput of stations having different traffic loads

TABLE III
TOTAL THROUGHPUT AND MAX-MIN FAIRNESS COMPARISON

Metric	CM-CSMA/CA	SRMC-CSMA/CA	HTFA
T (Mbit/sec)	41.20	47.6	49.3
F	0.31	0.07	0.05

backoff slot duration ranges from 20-120 μs. Fig. 13 shows the outcome where 10 stations actively participate. As we expected, the saturation throughput reduces as the backoff slot duration increases and vice-versa. Nevertheless, the OFDMA-employed multi-channel system still provides more throughput than the single-channel system up to a certain value of the backoff slot duration, for example, up to about 82 microseconds for M=3 sub-channels. As the number of sub-channels increases, the outperforming range of the backoff slot duration increases and vice-versa.

B. *Protocol Comparison*

Here, we use the symbol l_i to denote the traffic load and t_i to denote achieved throughput of the i^{th} terminal. Obviously, $t_i \leq l_i$. As traffic loads of different terminals may vary significantly, we wish to find normalized throughput (t_i/l_i) for a fair comparison. We use two different metrics to examine the performance efficiency of our designed protocol. These two metrics are total throughput of the network denoted by T and the max-min fairness denoted by F which are measured according to the following equations:

$$T = \sum_{i=1}^{N} t_i \tag{9}$$

$$F = \max\frac{t_i}{l_i} - \min\frac{t_i}{l_i} \tag{10}$$

Table II listed the simulation parameters for subsequent experiments. We assume terminals generate packets according to Poisson distribution and different terminals may have different traffic loads. We compare our HTFA protocol with the CM-CSMA/CA protocol proposed in [3] and SRMC-CSMA/CA protocol proposed in [4] in the following scenario.

We conduct simulation for three terminals having loads 12, 18 and 24 Megabits respectively and we referred those as low load, medium load and high load terminal. We assume total bandwidth of the whole channel is 54 Mbit/sec and the channel is evenly partitioned into three sub-channels. Thus, each of the sub-channels gains a bandwidth of 18 Mbit/s. The normalized throughput of all terminals (low, medium and high load) is measured for the intended three protocols and is shown in Fig. 14. Analyzing the CM-CSMA/CA protocol, we see the normalized throughput reduces significantly with the increases in the traffic load. In CM-CSMA/CA protocol, one terminal can access only one sub-channel in most cases and hence the normalized throughput of the high load terminal does not exceed 0.60. On the other hand, according to HTFA and SRMC-CSMA/CA protocol, any terminal could acquire multiple sub-channels and transmits data concurrently. Therefore, the normalized throughput of the high load terminal is not bounded by the bandwidth of an individual sub-channel. We also observe that the normalized throughput of all the three types of loads is above 0.81 of HTFA and SRMC-CSMA/CA protocol. However, HTFA performs slightly better than SRMC-CSMA/CA due to less contention and less collision ensured by HTFA.

The overall throughput and max-min fairness comparisons are summarized in Table III. The terminals in HTFA will not contend for sub-channel access rather than sub-channels are dedicated to the terminals if the number of terminals (N) is smaller or equal to the number of sub-channels (M). If $N \leq M$, terminals monopolize the sub-channel access and there is no random backoff slot. Thus, HTFA provides higher throughput than CM-CSMA/CA and SRMC-CSMA/CA. Our HTFA protocol employs a hybrid mechanism to distribute the sub-channels among the terminals. Thus, its max-min fairness is promising than CM-CSMA/CA and SRMC-CSMA/CA.

VII. CONCLUSION

The rapid growth of demand for high-speed WLAN has driven intense research to enhance the throughput by employing a variety of medium access control (MAC) mechanisms. The efficiency of MAC plays a major role to enhance the throughput of a Wireless LAN system. One of the promising access mechanisms for MAC is orthogonal frequency division multiple access (OFDMA). In this paper, we propose an OFDMA-based MAC protocol named 'HTFA' which employs a hybrid mechanism for channel access. HTFA will provide high throughput of data as well as maintains improved fair access policy to the medium among the terminals. The main distinguishing feature of our proposed protocol is its uniqueness in distributing the sub-channels to the terminals. We perform rigorous simulations with network simulator-3 that is presented in Section VI. Simulation results confirm validation of our protocol in terms of throughput, collision reduction and fairness. We get these advantages at the expense of increased complexity in the channel distribution procedure. Still, we convinced that there would be a tremendous tradeoff of the proposed protocol. Theoretical analysis of saturation throughput of the HTFA protocol is also evaluated in Section V employing an ideal comprehensive model.

Through simulation, we also able to approximate the number of sub-channels in accordance with the number of participating terminals in a wireless LAN where all sub-channels are equally dispersed. Throughout the article, we consider the sub-channels are of equal length to avoid complexity and to keep within the scope of the paper. However, it is also urged to establish a model and carry out simulations and mathematical analysis for varying sub-channel length which might be subject to future research.

REFERENCES

[1] Jing Wang, Guixia Kang and Ping Zhang, "A two-dimensional medium access control protocol based on OFDMA and CSMA/CA," Wireless Telecommunications Symposium (WTS), pp. 1-5, April 2011.

[2] Hojoong Kwon, Hanbyul Seo and Seonwook Kim, "Generalized CSMA/CA for OFDMA systems: protocol design, throughput analysis, and implmentation issues," IEEE Transactions on Wireless Communications, Vol. 8, No. 8, pp. 4176-5187, August 2009.

[3] Xudong Wang and Hao Wang, "A novel random access mechanism for OFDMA wireless networks," Global Telecommunications Conference (GLOBECOM), Dec. 2010.

[4] Jia Xu, Pin Lv and Xudong Wang, "Single-radio multi-subchannel random access for OFDMA wireless networks," Electronics Letters, Vol. 49, No. 24, pp. 1574–1576, November 2013.

[5] Qiao Qu, Bo Li, Mao Yang and Zhongjiang Yan, "An OFDMA based concurrent multiuser MAC for upcoming IEEE 802.11ax," IEEE Wireless Communications and Networking Conference Workshops, March 2015.

[6] Cao Xuelin, Song Zuxun and Yang Bo, "TR-MAC: A multi-step slot reservation-based hybrid MAC protocol for ad hoc networks," IEEE International Conference on Signal Processing, Communications and Computing (ICSPCC), pp. 1 - 5, Sept. 2015.

[7] Yunjie Yuan, Bo Li, Yi Chen and Hu Zhou, "CCRM: A MAC protocol with cooperative channel reservation for wireless ad hoc networks," 7th International Conference on Wireless Communications, Networking and Mobile Computing, pp. 1 - 4, Sept. 2011.

[8] Choi, Y. J., Park, S. and Bahk, S., "Multichannel random access in OFDMA wireless networks," IEEE Journal on Selected Areas in Comm. pp. 603–613 VOL. 24, NO. 3, March 2006.

[9] IEEE 802.11, Part 11: Wireless LAN medium access control (MAC) and physical layer (PHY) specifications, IEEE 802.11 Std., 2018.

[10] NS-3, "Network Simulator," https://www.nsnam.org/

[11] Md. Mustafizur Rahman, Choong Seon Hong and Sungwon Lee, "A high throughput on-demand routing protocol for multirate ad hoc wireless networks," IEICE Transactions on Communications, pp. 29-39 VOL. E93–B, NO.1 January 2010.

[12] G. Bianchi, "Performance analysis of the IEEE 802.11 distributed coordination function," IEEE J. Select. Areas in Commun., vol. 18, no. 3, pp. 535-547, March 2000.

[13] I. Tinnirello, G Bianchi and Y. Xiao, "Refinements on IEEE 802.11 distributed coordination function modeling approaches," IEEE Transactions on Vehicular Technology, vol. 59, no. 3, pp. 1055-1067, March 2010.

[14] IEEE 802.11, Proposed 802.11ax Functional Requirements, IEEE 802.11-14/0567r7, July 2014.

[15] IEEE 802.11, HEW MAC Efficiency Analysis for HEW SG, IEEE 802.11-13/0505r0, May 2013.

[16] IEEE 802.11, Uplink Multi-user MAC Protocol for 11ax, IEEE 802.11-14/0598r0, May 2014.

[17] IEEE 802.11, Discussion on OFDMA in IEEE 802.11ax, IEEE 802.11-14/0839r1, Jul. 2014.

[18] "Introduction to 802.11ax High-Efficiency Wireless," White paper of 'National Instruments', June 2016. http://www.ni.com/white-paper/53150/en/

[19] Ian F. Akyildiz and Xudong Wang, "Wireless Mesh Networks" Wiley, pp. 52-53, 2009.

[20] G. Haile and J. Lim, "C-OFDMA: Improved throughput for next generation WLAN systems based on OFDMA and CSMA/CA," 4th International Conference on Intelligent Systems, Modelling and Simulation, pp. 497-502, January 2013.

[21] H. Ferdous and M. Murshed, "Enhanced IEEE 802.11 by integrating multiuser dynamic OFDMA," Wireless Telecommunications Symposium (WTS), pp. 1–6, April 2010.

[22] Gazi Zahirul Islam and Mohammod Abul Kashem, "An OFDMA-based new MAC mechanism for IEEE 802.11ax," 5th International Conference on Networking, Systems and Security (NSysS), pp. 1-7, December 2018.

Gazi Zahirul Islam is currently pursuing his PhD at Bangladesh University of Professionals. He completed M.Sc. in wireless communications systems engineering from the University of Greenwich, UK and B.Sc. in computer science and engineering from Chittagong University of Engineering and Technology (CUET), Bangladesh. He has been teaching as an Assistant Professor at Daffodil International University, Bangladesh for about 5 years. Previously he taught at City University, Southern University Bangladesh and Primeasia Univesity. He is an author of several research articles on wireless communications and e-governance.

Mohammod Abul Kashem is a Professor of Computer Science and Engineering Department of Dhaka University of Engineering and Technology (DUET), Bangladesh. He also served as the director of "Institute of Information and Communication Technology" of DUET. He is a Post Doctorate of University Lumiera Lyon2, France and he completed PhD from National University "Lviv Politechnic" Ukraine. He also passed M.Sc. Engg. and B.Sc. Engg. from State University "Lvivska Polytechnica" Ukraine. His fields of specializations are mobile communication, information system management, cyber physical system and speech signal processing. He is an author of more than 50 articles on his research areas.

R3D3: A Doubly Opportunistic Data Structure for Compressing and Indexing Massive Data

Máté Nagy, János Tapolcai and Gábor Rétvári

***Abstract*—Opportunistic data structures are used extensively in big data practice to break down the massive storage space requirements of processing large volumes of information. A data structure is called (singly) opportunistic if it takes advantage of the redundancy in the input in order to store it in information-theoretically minimum space. Yet, efficient data processing requires a separate index alongside the data, whose size often substantially exceeds that of the compressed information. In this paper, we introduce doubly opportunistic data structures to not only attain best possible compression on the input data but also on the index. We present R3D3 that encodes a bitvector of length n and Shannon entropy H_0 to nH_0 bits and the accompanying index to $nH_0(1/2 + O(\log C/C))$ bits, thus attaining provably minimum space (up to small error terms) on both the data and the index, and supports a rich set of queries to arbitrary position in the compressed bitvector in $O(C)$ time when $C = o(\log n)$. Our R3D3 prototype attains several times space reduction beyond known compression techniques on a wide range of synthetic and real data sets, while it supports operations on the compressed data at comparable speed.**

***Index Terms*—succinct and compressed data structures, compressed self-indexes, big data, packet forwarding**

I. INTRODUCTION

Recently, the exponential growth of available electronic information has created new challenges in data mining, machine learning, pattern analysis, and networking, as the sheer volume of data to be stored, transferred, and processed online has greatly surpassed the increase in memory, disk, and link capacities of current computers and computer networks [1], [2]. Space reduction for massive data processing applications is an attractive choice to tackle these challenges, as storage space is fundamentally related to the time it takes to process data [3]. In fact, by making better use of cache and memory levels closer to the processor, waiving the painful cost of disk accesses, and utilizing processor–memory bandwidth more efficiently, space reduction techniques can make processing of unprecedentedly large quantities of data feasible even in resource-constrained environments. Ultimately, the goal is to *store data in memory in a compact or compressed format* and *still operate directly on it* without any major performance hit compared to a naive, uncompressed representation [4].

Succinct and compressed data structures are a relatively new development in theoretical computer science that promise with substantial decrease in the memory footprint of big data operations, by storing sequential or structured static data in a compressed but readily accessible, queryable, and manipulable format [5]. Applications encompass essentially the entire field of computer science, from space-efficient encodings of ordered sets, sparse bitmaps, partial sums, binary relations, range queries, and arbitrary sets supporting predecessor and successor search [6]–[9], ordinal and labeled trees [6], [9]–[13] and general graphs [6], [14], indexing massive textual data [5], [12], [15]–[19], top-k document retrieval, suffix trees, arrays, and inverted indexes in information retrieval systems [12], [16], [19]–[22], point grid queries in computational geometry [22] and genome compression in computational biology [23], [24], all the way to key-value stores, log analytics, machine learning, data mining, and big data applications [4], [25].

The cornerstone of these schemes is a *compressed bitvector representation* that encodes an arbitrary bitmap in very small space and, at the same time, implements some simple operations, namely access, rank, and select queries (see later), right on this compactified format [6], [15], [17], [26]–[30]. Such compressed bitvectors can then be used to build composite data structures and construct complex queries on them [6]. As recently shown, for instance, such compressed bitvectors can be used to construct a space-efficient representation for Internet routers' forwarding tables (FIBs) [31]. The resultant compressed IP FIBs have been shown to squeeze the routing table of a contemporary IPv4 router, counting beyond 500,000 prefixes, to a mere 70–200 kbytes of memory, while supporting wire-speed longest-prefix matching right on the compressed form.

Space usage of any queryable data structure boils down to two elementary components: the space for storing the data itself, plus some additional space for an index into the data that guarantees fast access [25]. In this setting, the data component constitutes the useful information and the index is pure redundancy, whose size should be minimized as much as possible. The first technique to attain worst-case-optimal storage space on both the data and the index components was the *succinct bitvector and ordered-tree data structures* due to Jacobson [6] (but see also [15]). The memory footprint was further reduced by Ferragina and Manzini, who introduced *opportunistic data structures that attain information-theoretically minimal entropy-constrained storage space on the data component* [20]. Their data structures are called (singly) opportunistic in that they can take advantage of the compressibility of the input by decreasing the space occupancy beyond the worst-case limit, at no significant slowdown in query performance. A good example for such an opportunistic compression approach is the RRR compressed bitvector scheme due to Raman, Raman, and Rao [12], attaining nH_0 bits on the data and $O(\frac{n \log\log n}{\log n}) = o(n)$ bits on the index, where n is the length of the input and H_0 is the zero-order

Máté Nagy, János Tapolcai and Gábor Rétvári are with MTA-BME Lendület Future Internet Research Group, MTA-BME Information Systems Research Group Department of Telecommunications and Media Informatics, BME, Budapest, Hungary (email: {nagym, tapolcai, retvari}@tmit.bme.hu)

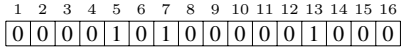


Figure 1: A sample bitvector.

empirical entropy [32], while supporting access, rank, and select queries in optimal $O(1)$ time. Today, the RRR scheme serves as the major building block for space-efficient data processing techniques, enjoying wide-scale use throughout the entire spectrum of compressed information processing [12], [16], [18], [19], [27], [29].

A major shortcoming of compressed information processing is, however, that the storage size of the index can significantly outweigh (up to and beyond 8 times, [4]) that of the data, taking a huge toll on the storage efficiency of data compression and hindering engineering applications [27]–[30]. To address this limitation, in this paper we introduce the concept of *doubly opportunistic data structures*, which, as opposed to conventional opportunistic schemes that compress only the data component, *achieve information-theoretically minimal entropy-constrained space both on the data and the index at the same time*. We present R3D3 ("RRR–Developed Data structure for big Data"), which combines the storage scheme of RRR for encoding the index and the Elias-Fano compression method [15] for block-encoding the data, to attain $nH_0 + nH_0(\frac{1}{2} + O(\frac{\log C}{C}))$ bits of space and random access and rank queries in $O(C)$ time and select in $O(\log n)$ when $C = o(\log n)$ constant. R3D3 thusly not only attains provably maximum compression (up to small error terms) on both the data and the index, and hence qualifies as the first doubly-opportunistic bitvector compression scheme, but it also allows to realize many interesting engineering trade-offs between storage space and query time by fine-tuning the constant C. By comprehensive evaluations on synthetic data sets and a real data corpus we show that R3D3 achieves from 2 up to 10 times smaller space than RRR while supporting queries in similar, or slightly worse, performance.

The rest of the paper is structured as follows. In Section II we review bitvector compression, in Section III we introduce R3D3 and give a detailed space–time analysis, in Section IV we present the results of our benchmarks, and finally we conclude our work in Section V.

II. COMPRESSED BITVECTOR INDEXING

In this section we give an overview on succinct and compressed data structures and we describe the RRR and the Elias-Fano coding schemes in some detail.

A. Notations and Definitions

Let t be a bitvector with length n. The number of bits set to 1 in t is called the population (or popcount) and the ratio of the population and n is the empirical probability p of 1s in t. Our aim is to build a compact representation for t that supports the following queries efficiently:

- $\mathrm{access}(t, i)$: return the i-th bit of t;
- $\mathrm{rank}_q(t, i)$: return the number of occurrences of symbol q in $t[1, i]$;
- $\mathrm{select}_q(t, i)$: return the position of the i-th occurrence of symbol q in t.

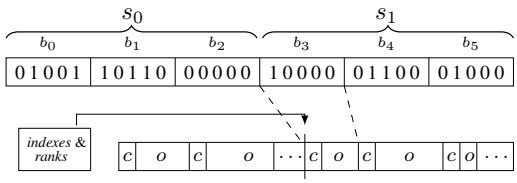


Figure 2: Sketch of the RRR encoding scheme.

Consider the example in Fig. 1. Here, $n = 16$, $p = 3/16$ and $\mathrm{popcount}(t) = 3$, the query $access(t, 5) = 1$ tells that the bit at position 5 is set, $rank_1(t, 8) = 2$ gives the number of bits set to 1 up to and counting the 8-th position, and finally $select_1(t, 2) = 7$ indicates that the second set bit occurs at position 7. Notice that rank and select are "dual" in that if $\mathrm{select}_1(t, i) = m$ then $\mathrm{rank}_1(t, m) = i$. Further, $\mathrm{rank}_0(t, i) + \mathrm{rank}_1(t, i) = i$ but the same does not hold for select.

A *succinct encoding* of t will store t on worst-case minimum $n+o(n)$ bits of space (the uncompressed representation would need n bits and the error term $o(n)$ vanishes asymptotically) and implement access, rank, and select "fast" (preferably in $O(1)$). The naive "bitmap" representation is not succinct in this sense since it fails the second requirement; rank and select would need a linear sweep through the bitmap, taking $O(n)$ time. A *compressed encoding* of t, on the other hand, reduces the memory footprint beyond the worst-case limit, if the input is compressible, to $nH_0+o(n)$ bits, where H_0 is the zero-order empirical entropy (or the Shannon entropy) of t:

$$H_0 = p \log \frac{1}{p} + (1-p) \log \frac{1}{1-p} \leq 1 \ ,$$

without any performance penalty on the performance of queries. Note that all our logarithms are base 2. For brevity's sake, we shall mostly omit rounding our logarithms to integers in the forthcoming analyses wherever this does not affect the validity of the results.

B. A Scheme due to Raman, Raman, and Rao

Raman, Raman, and Rao introduced the first compressed data structure for bitmaps, usually referred to as *RRR*, that solves access and rank queries in constant time [12]. In this paper, we describe a modified encoding due to Navarro and Providel [30], which, although needs slightly worse $O(\log n)$ time for queries, proved much more space- and time-efficient in practical implementations [33].

RRR comprises a block-coding component to encode the useful data and an indexing scheme to support queries to the blocks [27]–[30]. The structure partitions t into blocks $b_1, b_2, \ldots$ of size $b = \frac{\log n}{2}$ bits (see Fig. 2 for an illustration). Each block b_i is encoded with a pair (c_i, o_i), where $c_i = \mathrm{popcount}(b_i)$ is the *class* of b_i and o_i is the offset, or the *combinatorial rank*, of b_i, defined as the sequence number of b_i in some fixed enumeration (e.g., lexicographic order) of all combinations of exactly c_i occurrences of 1s on b bit positions [34]. Storing c_i needs $\log(c_i + 1)$ bits and o_i needs $\log \binom{b}{c_i}$, so the block codes (the *data component*) take $\sum_i \log(c_i + 1) + \log \binom{b}{c_i} = nH_0 + O(\frac{n}{\log n})$ bits overall [35].

The *indexing scheme* in turn groups every consecutive $\log n$ blocks into a superblock. Then, for each superblock the index stores the starting positions for the block-codes inside it and the cumulative rank up to the superblock's beginning, plus, for each block, the corresponding block-code's starting position and the rank at the block's beginning, both relative to the superblock that contains it. Cumulatively, this indexing structure needs $O(\frac{n \log\log n}{\log n}) = o(n)$ bits of space.

Answering $\mathrm{access}(t, i)$ works as follows. As superblocks and blocks span constant number of bits in t, i uniquely determines the superblock and the block that contain position i. We follow first the superblock pointer and then the block pointer to reach the block-code for the corresponding position, this can be done in $O(1)$ time. From this point, decoding a block (the so called *combinatorial unranking* operation) takes $O(b) = O(\log n)$ time [30], [34], [36]. Solving rank goes similarly, but this time we also add up the superblock's and block's rank counters along the way, which, together with the time to unrank the block, takes $O(\log n)$ time. Finally, select binary-searches over superblock and block ranks, again in $O(\log n)$ time.

Experimental studies show that the $O(\frac{n \log\log n}{\log n})$ bits size of the index, although asymptotically small, may outweigh the data components' size nH_0 substantially, especially for low-entropy input [27]–[30]. Correspondingly, many schemes eliminate block-code pointers and rank counters from inside the superblocks, which tends to save a lot of space at the cost of degrading block access and rank to a linear search over the blocks of the superblock, making queries slow. This scheme is usually referred to as, somewhat confusingly, the *unindexed* version of RRR, to distinguish it from the above described version (with explicit block pointers and ranks inside superblocks) that is called *indexed* RRR.

Today, RRR is a popular tool amongst theoreticians and practitioners and constitutes a fundamental building block for compressed indexes of complex structured and unstructured types of information, like trees [13], strings (wavelet trees, [16]), or IP forwarding tables [31]. Practice has shown, nevertheless, that RRR exhibits a brittle space–time trade-off: meaningful storage space reduction can only be realized at the price of sacrificing precious query performance, like adopting larger block sizes [30] or swapping indexed-RRR to the much slower unindexed version [33].

C. The Elias-Fano scheme

Elias-Fano coding has been proposed in [15] to store a bitvector t in $nH_0 + o(n)$ space and answer select_1 queries $O(1)$ time, with no support for rank and access. Herein, we describe an alternative scheme *EF* that attains $nH_0 + o(m)$ bits of space and needs $O(m)$ for access, select, and rank, where $m = \mathrm{popcount}(t)$ (see also [12], [22], [25], [27], [37]).

The idea of EF is to encode the characteristic vector $\{x_1, x_2, \ldots, x_m\}$ of t, where $x_i = select_1(t, i) : i \in \{1, \ldots, m\}$, instead of t itself. EF uses a technique called MSB bucketing: group x_is according to the most significant $\log m$ bits into buckets, store the $l = \log n - \log m = \log \frac{n}{m}$ lower-order bits for each x_i verbatim in an array (called the Lower-bits Array, *LBA*), and store the significant bits as a sequence of unary encoded gaps in another array (the Upper-bits Array, *UBA*) as follows: for each bucket write down as many 1s as there are x_is in the bucket followed by a 0.

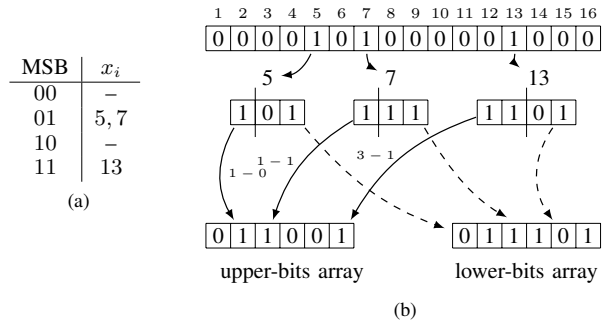


Figure 3: Elias-Fano encoding scheme: (a) MSB bucketing on the characteristic vector $(5, 7, 13)$, and (b) EF encoding.

Perhaps an example is in order here. In Fig. 3, $x_1 = 5$, $x_2 = 7$ and $x_3 = 13$, $n = 16$ and $m = 3$, so $l = \lfloor \log 16/3 \rfloor = 2$. This means that the LBA will contain the lower $l = 2$ bits of each x_i verbatim. Further, the bucket size is $2^l = 4$, and so the number of x_is in each bucket is $0, 2, 0, 1$, whose unary encoding gives the UBA: 0110010 (the last 0 can be omitted).

Storing the m elements of the LBA takes $m \log \frac{n}{m}$ bits while the UBA needs $2^{\log m} + m = O(m)$ bits, as there are as many 0s as there are buckets plus m bits set to 1 for each x_i in $i \in \{1, \ldots, m\}$. Finally, we need an additional $\log m$ bits to store m, which we omit here for reasons that will be made clear later. The overall size of EF is $m \log \frac{n}{m} + 2^{\log m} + m = nH_0 + O(m)$ bits, where the data component (the LBA) takes nH_0 bits and the index (the UBA) takes another $O(m)$.

Now, answering $\mathrm{access}(t, i)$ goes as follows. First, we find the bucket q that contains position i: $q = \frac{i}{2^l}$, then we find the run of 1s in the UBA that corresponds to the q-th bucket: $z = \mathrm{select}_0(UBA, q)$; we observe that there were exactly $z - q$ occurrences of 1s in the UBA *before* position z so we scan the LBA *leftward* from position $(z - q)l$, decoding at most 2^l elements x_j of the characteristic vector; if for some $x_j = i$ then the result of the query is 1, otherwise 0. For instance, $\mathrm{access}(t, 6) = 0$ in Fig. 3, as position 6 is in the second bucket thus the MSB is 01, $q = \mathrm{select}_0(UBA, 2) = 4$ so up until the end of the second bucket there were $4 - 2 = 2$ bits set to 1, and decoding the LBA from the second entry leftward, combined with the MSB 01, yields first 7 and then 5, at which point we know the answer is 0. This goes in $O(m)$ time, as just answering the first select query may require a linear search on the UBA in the worst case. Note that adding another $O(m)$ bits would guarantee $O(1)$ random access [22], [37], but we disregard this option here as it would double the index size. Solving rank goes similarly, while select is by binary search over the UBA and the LBA, again in $O(m)$ time.

When compared to RRR, EF usually yields larger encoded size. At the extreme, for $p = 0.5$ EF uses $1.5n$ bits, a whopping 50% overhead. Furthermore, the somewhat rigid structure of EF does not provide too much in the way of the space–time trade-off like the one we have seen for RRR. Then again, EF can be very fast depending on the input t, as queries take only $O(\mathrm{popcount}(t))$ steps; this can be a massive win, e.g., for small-entropy input. Our compressed bitvector data structure, R3D3 to be presented next, heavily builds on this property.

III. A DOUBLY OPPORTUNISTIC DATA STRUCTURE

In summary, both RRR and EF are opportunistic data structures that realize significant space savings in the data encoding, with EF yielding potentially faster but larger encodings than RRR. Could we somehow combine RRR and EF into a compressed bitvector scheme that would somehow display the advantages of both simultaneously?

In this section we answer this question in the affirmative. We propose R3D3, a combination of RRR and EF that, in contrast to conventional singly-opportunistic encodings that compress only the data component, attains entropy-constrained size on both the data and the index. Thus, we call R3D3 a *doubly opportunistic* data structure.

A. R3D3

So how can we combine the advantages of RRR and EF? First, RRR's indexing scheme gives very fast $O(1)$ access to block-codes and block-ranks, so we definitely want to keep it. It also offers an elegant way to tune the space–time trade-off: The RRR index size is chiefly shaped by the block size b; the larger the block size the fewer blocks we need, and hence the fewer the costly block pointers and block-ranks. Since these dominate the size (taking $O(n/\log n)$ bits when $b = \log n$), increasing blocks will go to great lengths to save memory on indexing. Unfortunately, this cannot be done with RRR for free, as the access and rank execution times are dominated by the block-coding component's running time $O(b)$.

But what if we substitute the block-coding component with EF? After all, decoding a block b_i requires only $O(c_i)$ steps with EF where, recall, c_i is the class of b_i: $c_i = \mathrm{popcount}(b_i)$, in contrast to the $O(b)$ time complexity of combinatorial unranking; in other words, *EF's efficiency depends fundamentally on the number of* 1*s in a block and not the block size itself*. Hence, we can safely increase the block size b to save space on RRR's indexing until we reach block-coding execution-time parity with RRR, which occurs when $\mathrm{popcount}(b_i) = O(\log n)$. At this point our larger blocks will contain as many 1s as the default block size $O(\log n)$ of RRR and so both will need $O(\log n)$ steps for block-decoding, but we gain significant space on the indexing, thanks to the large blocks. Then again, EF-coded blocks will be slightly larger than in RRR, but the gross space reduction we earn on the index will, hopefully, amply compensate for this loss.

This is the main idea of R3D3: we keep the indexing structure of RRR but we swap the block-coding component for the much more efficient EF. Then, we can increase blocks way beyond what RRR would admit, without major penalty on query times. The basic structure of R3D3 is given in Fig. 4.

Building the R3D3 encoding goes very similarly to how it happens with RRR, just the block-coder is now EF instead of combinatorial ranking/unranking. First, we divide the input t into superblocks of size s and blocks of size b (we set these parameters later), build the RRR index, encode the class c_i for each block b_i directly and then invoke EF to encode b_i. Note that the input to EF is now the block b_i and the length equals b. Additionally, the number of 1s in b_i (the input parameter m) is exactly c_i, so we do not need to store it separately in EF. To control c_i and get better compression we do the usual trick that if $\mathrm{popcount}(n) > \frac{n}{2}$ then we encode the inverse of t instead of t. In fact, in our implementation we do this trick block-wise [33], which yields $p < \frac{1}{5}$ and $c_i < \frac{b}{5}$.

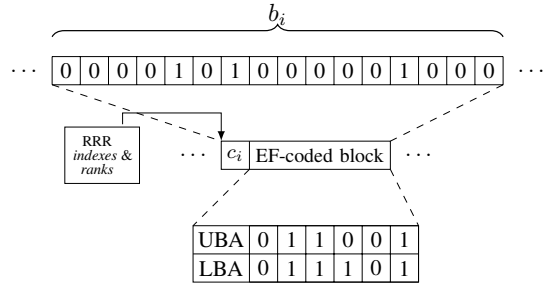


Figure 4: Sketch of the R3D3 encoding scheme, with a single 16-bit block and the corresponding EF block-code. The pointers and rank counters and the block classes are encoded in the RRR index, while the blocks are encoded with EF.

In fact, R3D3 adopts a scheme we call *duplicate indexing*; it first invokes the RRR indexes to find the starting position for each block, then looks up the UBA to index the relevant entries in the LBA, and finally only a few LBA entries need to be directly decoded. As the analysis in next section reveals, this duplicate indexing scheme yields a highly space- and time-efficient compressed bitvector data structure.

B. Analysis

We fix the superblock size at $s = b \log n$, like in RRR (see the proofs in the Appendix for the reason); the block size b will be determined later. With this parameter setting, the result below gives the storage space and the query times for R3D3.

Theorem 1. *Let t be a bitvector of length n, let $p = \mathrm{popcount}(t)/n$, let H_0 be the zero-order empirical entropy of t, and fix the block size at b. Then, encoding t with R3D3 needs at most*

$$nH_0 + np + \frac{n}{b}\left(2 + 3\log b + 2\log\log n\right)$$

bits and supports access *and* rank *queries in expected $O(pb)$ time and* select *queries in $O(\log n)$ if $pb = o(\log n)$.*

We give the proof of Theorem 1 through a sequence of technical Lemmas; for clarity the proofs of the Lemmas in turn will be relegated to the Appendix.

The below Lemma characterizes the encoded size M_I of the RRR index structure that we embed into R3D3.

Lemma 1. *The RRR index needs $M_I = \frac{n}{b}(2 + 3\log b + 2\log\log n)$ bits.*

M_I is of course the redundancy in R3D3. Next, we give the size of the EF-coded blocks, M_D.

Lemma 2. *The EF-encoded data needs $M_D = nH_0 + np$ bits.*

Finally, the query execution times stated below for R3D3 are as follows: for $\mathrm{access}(t, i)$ locating the beginning of the EF-coded block that contains position i and identifying the class take $O(1)$ time, to which block-decoding adds another $O(pb)$ for the "average" block. The same holds for $\mathrm{rank}(t, i)$, while $\mathrm{select}(t, i)$ goes with binary-searching superblock and block ranks in $O(\log n)$ time and then decoding the block, again in expected $O(pb)$ time. The total time $O(\log n) + O(pb)$ is dominated by the binary-search as long as $pb = o(\log n)$.

Lemma 3. *Answering* access *and* rank *queries on the R3D3 representation needs expected* $O(pb)$ *time and* select *goes in* $O(\log n)$ *as long as* $pb = o(\log n)$.

This completes the proof of Theorem 1. What remains to be done is to fine-tune the block size b. This needs to be done very carefully; increasing b makes for smaller index M_I and hence smaller overall size (the data part M_D is by and large independent of b), but increasing b too much deteriorates query time. We need to strike a fine balance between space and time here, one that results in entropy-constrained size for both M_D and M_I but still does not ruin query performance.

We introduce a new parameter $C = pb$, which can be broadly interpreted as the "average" popcount of blocks. Of course, $C \geq 1$ to ensure that there is at least one bit set in each block. Thus, $b = C/p$ and we immediately get the execution times for access and rank as $O(pb) = O(C)$. Then again, C must not be too large, that is, beyond $O(\log n)$, otherwise select suffers. This gives the useful range $C \geq 1$, $C = o(\log n)$. The following result summarizes these findings.

Theorem 2. *Let* t *be a bitvector of length* n *and entropy* H_0*, and let* $C \geq 1$*,* $c = o(\log n)$*. Then, encoding* t *with R3D3 needs at most*

$$nH_0 + nH_0\left(\frac{1}{2} + O\left(\frac{\log C}{C}\right)\right) \tag{1}$$

bits and supports access *and* rank *in expected* $O(C)$ *time and* select *in* $O(\log n)$.

Again, consult the Appendix for the proof.

C. Discussion

We close this Section with some remarks on R3D3.

First, R3D3 achieves entropy-constrained space on both the data and the index (up to a small error term for the index): the RRR index and the UBA components in the block-codes, which, as per duplicate indexing, together make up the R3D3 index, need $nH_0\left(1/2 + O\left(\log C/C\right)\right)$ bits of storage space, while the data component (the LBAs) uses another nH_0 bits. As far as we are aware of, R3D3 is the first such doubly opportunistic compressed data structure.

Second, the above space bounds are strictly of worst-case nature, in that there are much tighter upper bounds than what we used in Theorem 2. Since $nH_0 + np \ll \frac{3}{2}nH_0$ when p is sufficiently small, the space bounds can be improved to $nH_0 + nH_0 O\left(\log C/C\right)$ bits if $p < 0.169$, a substantially tighter space characterization for low-entropy input.

Third, tuning constant C opens the door to a wide spectrum of space–time trade-offs. At one extreme, when $C = 1$, i.e., when there is only a single bit set per block on average, we get very fast $O(1)$ access and rank at the cost of a somewhat largish $nH_0\left(3/2 + O(1)\right)$ bits memory footprint, an overhead of $\sim 50\%$. This is because EF-coded blocks are slightly larger than RRR's blocks. On the other hand, increasing C will result larger blocks and less overhead for indexing; when $C = O(\log n)$ we get execution-time parity with RRR with much smaller $nH_0(\frac{1}{2} + O\left(\log\log n/\log n\right)$ bits indexes.

Finally, we observe that our results are in line with the lower bounds of [26], stating that we need $\Omega(\frac{\log\log n}{\log n})$ bits index to implement rank in $O(1)$. R3D3, however, gives $O(1)$ index size in this setting.

IV. NUMERICAL EVALUATIONS

Next, we turn to present a comprehensive set of experimental results to evaluate the space- and time-efficiency of R3D3. For this purpose, we created a proof-of-concept prototype on top of the Succinct Data Structure Library (SDSL, [33]), a powerful C++ template toolkit with comprehensive support for the state-of-the-art in compressed data structures. Stock SDSL offers only the unindexed version of RRR, therefore we created 3 additional C++ template classes on top of SDSL: indexed RRR plus indexed and unindexed versions of R3D3. In the rest of this section RRR and R3D3 will refer to the indexed versions. The R3D3 block-coding routines, furthermore, use the EF optimizations as described in [37]. The code is available at [38].

The two dimensions of interest are the compressed size and performance of queries for RRR and R3D3. We used the CPU's `RDTSC` register, holding the actual snapshot of the program counter, to measure execution times with (close-to) cycle-level precision. The experiments were conducted on a Linux PC, Intel Core i3 CPU @ 3.3GHz with 4Gbyte of RAM.

Block-coding. The goal of our first experiment is to validate our choice for EF instead of RRR's combinatorial ranking/unranking scheme to encode blocks. Recall, this choice was made because EF supports all basic block-operations in $O(\text{popcount}(b))$ time as opposed to $O(b)$ for RRR, where b is the block size, at the cost of slightly bigger block-codes. Note that the population of the block does not alter the relation between EF and combinatorial encodings.

We made 1 million trials, each time with a new random block generated by setting each bit to 1 independently with probability $p = 0.1$. Fig. 5 gives the average time to make a random access to the encoded blocks and Fig. 6 compares the average size of block-codes, as the block size increases from 16 to 128.

We observe that EF block-coding is indeed much less sensitive to the block size; while R3D3 needs only 3 times as much time to access a 128-bit block as for a 16-bit block, this factor is 25-fold with RRR. Furthermore, R3D3 produces only slightly larger blocks than RRR and both are comfortably close to the entropy bound (that RRR beats the entropy limit is not surprising, as combinatorial ranks are a maximally space-efficient universal code, plus our results do not account for the storage size of the class bits c_i). This seems a price it is well worth paying for more efficient block-(de)coding at higher block sizes as the reduced indexes will greatly compensate for this loss, as revealed in our next experiments.

Random synthetic bitmaps. For this experiment we stay at random bitmaps as input, but now we evaluate RRR and R3D3 *en bloc*, not just the block-coding components as previously. We generated 1 Mbit random bitmaps with increasing p from 0 to $1/2$ and we evaluated space and time characteristics of our compressed bitvectors; Fig. 7 gives the size and Fig. 8, Fig. 9, and Fig. 10 give the execution time for access, rank, and respectively select queries to random positions, averaged over 10 trials. We repeated the experiments for R3D3 at different settings for the block size: $b = 32$, $b = 64$, and $b = 256$, while for RRR we used the default setting $b = 16$.

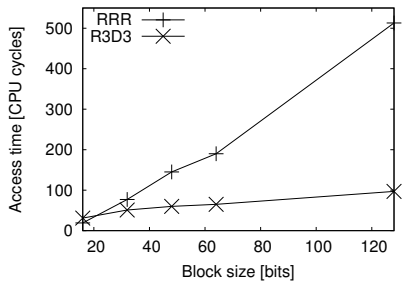


Figure 5: Average time to access a random position in RRR and EF block-codes as the function of the block size, on random bitmaps, $p = 0.1$.

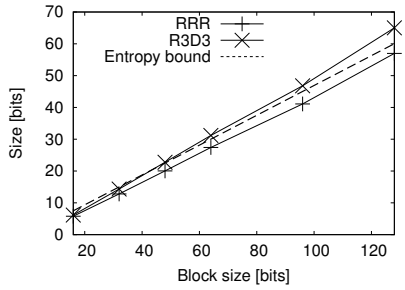


Figure 6: Average size of RRR and EF block-codes and the zero-order entropy limit (dashed line) as the function of the block size, on random bitmaps, $p = 0.1$.

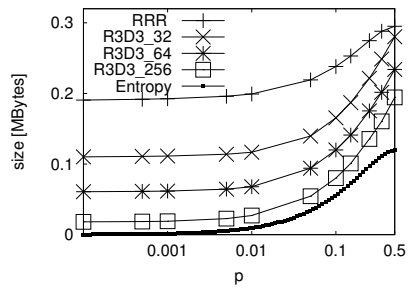


Figure 7: Average size of random bitmaps compressed with RRR and R3D3, and the corresponding entropy limit, as the function of p.

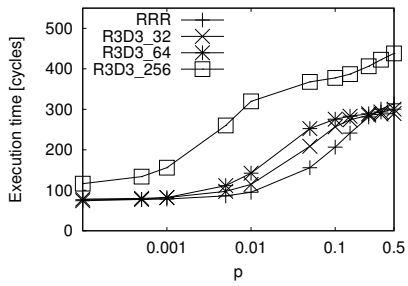


Figure 8: Average time of a random access query on random bitmaps.

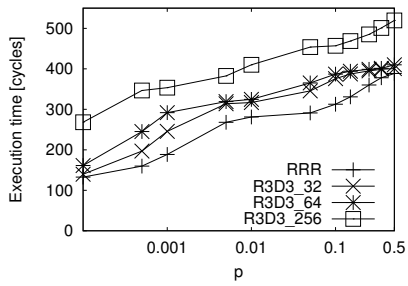


Figure 9: Average time of a random rank query on random bitmaps.

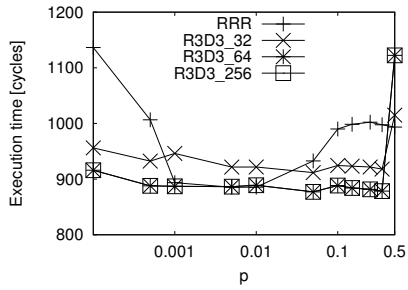


Figure 10: Average time of a random select query on random bitmaps.

On the storage size front, R3D3 exhibits huge gains over RRR. Even at $b = 32$ we already see two-fold reduction, while the setting $b = 64$ yields fourfold and $b = 256$ a whopping 4–10-fold improvement. At this point, R3D3 compresses very close to the entropy limit. On the other hand, the performance figures are slightly worse with R3D3; access is at most 20% and rank is at most 23% faster with RRR than with R3D3 when the block size is 32, the figures are 30% for access and 35% for rank at $b = 64$, and 30–70% on access and 10–60% on rank at $b = 256$. The performance difference manifests itself only on a limited regime of inputs and in the majority of the examined cases RRR and R3D3 produced remarkably similar performance figures. Finally select times are slightly better with R3D3, especially at larger block sizes.

Real data. We repeated the previous experiment, but this time over real data taken from real-life applications. For the first experiment we collected *bitmaps* from various sources of everyday engineering practice:

- `fax`: 1728x2376 bitmap image of text and diagrams from the Calgary Corpus [39];
- `bmp1`, `bmp2`: bilevel bitmap images scanned at 600dpi;
- `zip`: US ZIP codes in bitmap format, 1 marks a valid and 0 marks an invalid ZIP code;
- `caida_4`, `caida_8`, `caida_16`: adjacency matrices of the 4, 8, and 16-core of the Internet AS-level map in bitmap format, as obtained from CAIDA on 2014-06-01.

Table I: Comparison of RRR or R3D3 on real bitmaps: sample name, size, and entropy bound (nH_0); and compressed size and average execution time of random access and rank queries. Sizes are in Mbytes (MiB) and times in number of CPU cycles.

			RRR			R3D3_32			R3D3_64			R3D3_256		
Name	Size	Entropy	Size	Access	Rank	Size	Access	Rank	Size	Access	Rank	Size	Access	Rank
`fax`	0.49	0.19	0.86	106	205	0.56	112	256	0.37	125	267	0.2	210	328
`bmp1`	4.15	1.16	7.1	90	142	4.40	96	160	2.68	97	176	1.23	149	238
`bmp2`	4.15	1.69	7.46	121	223	4.99	141	278	3.34	150	290	2.06	236	357
`zip`	0.011	0.011	0.025	223	300	0.023	240	336	0.019	265	348	0.017	365	456
`caida_4`	3.38	0.19	5.6	82	145	3.28	87	195	1.84	95	243	0.62	146	328
`caida_8`	1.08	0.14	1.81	89	176	1.08	97	238	0.63	102	279	0.25	172	345
`caida_16`	0.34	0.09	0.58	104	211	0.37	114	282	0.23	124	297	0.11	209	354

Table II: Comparison of RRR or R3D3 on real textual data: sample name, size, and entropy bound (nH_0); and compressed size and average execution time of random access and rank queries. Sizes are in Mbytes (MiB) and times in number of CPU cycles.

			RRR			R3D3_32			R3D3_64			R3D3_256		
Name	Size	Entropy	Size	Access	Rank	Size	Access	Rank	Size	Access	Rank	Size	Access	Rank
shakes	0.119	0.068	0.18	3860	1495	0.17	3402	1614	0.14	3473	1639	0.115	4477	1802
scifi	0.733	0.401	1.04	3619	1523	0.97	3219	1625	0.79	3280	1658	0.65	4201	1809
bible	3.86	2.06	5.26	3451	1504	4.83	3094	1632	3.99	3156	1655	3.26	4021	1806
chr7	10	2.5	6.61	1518	771	5.9	1427	825	4.79	1441	836	3.88	1840	915
chr22	3.73	0.92	2.45	1523	778	2.17	1427	829	1.77	1433	841	1.42	1825	913
coli	4.42	1.11	2.75	1482	769	2.57	1379	822	2.14	1403	839	1.77	1813	923
euler	1.91	0.79	2	2623	1146	1.83	2378	1210	1.51	2404	1238	1.23	3077	1368
pi_1M	0.95	0.39	1.02	2640	1159	0.93	2401	1208	0.76	2430	1241	0.62	3107	1360
pi_10M	9.54	3.96	10.23	2643	1161	9.28	2403	1229	7.62	2429	1242	6.19	3103	1357

Table III: Comparison of RRR or R3D3 on routing tables: sample name, number of prefixes, and entropy bound as of [31]; and compressed size and average execution time of random FIB lookups. Sizes are in Kbytes (KiB) and times in CPU cycles.

			RRR		R3D3_32		R3D3_64		R3D3_256	
Name	#Prefixes	Entropy	Size	Lookup	Size	Lookup	Size	Lookup	Size	Lookup
hbone-szeged	453,685	70.1	172.7	11468	145.8	9092	116.7	10444	93.2	14724
access_d	403,245	149.1	226.1	10828	193.7	9576	155.4	10268	123.8	13492
access_v	2,970	1.08	7.6	5672	7.4	5448	6.6	6772	6.4	7796
mobile	4,391	1.32	3.7	6760	3.8	6488	3.6	7260	3.5	7588
hbone-vh1	453,741	222.6	418.5	9248	362	7600	293.7	7556	238.1	9264

The results are given in Table I. The first surprising observation is that not just that RRR does not reach the entropy limit but it completely fails even the uncompressed size. This is due to the excessive size of the index that we need to store to allow queries into the compressed data. R3D3, on the other hand, attains at least the uncompressed size at $b = 32$, improving on RRR by a factor of 2 in most cases. Increasing the block size to 64 then decreases the size by another factor of 2, while at $b = 256$ R3D3 gets very close to the entropy limit, improving over RRR by around a factor of 8. Meanwhile, the performance of queries with R3D3 remains comfortably close to that for RRR: at $b = 32$ the access execution times are on par and rank is at most 30–40% slower, while at $b = 256$ we get roughly half the performance of RRR. Recall, this is in return to about 8 times smaller size.

We repeated the experiments with *textual data*, this time compressing the input using Huffman-shaped wavelet trees [16]. Since a wavelet tree is essentially just a collection of bitmaps organized into a tree structure and access and rank queries translate to those on these bitmaps, wavelet trees nicely exercise the underlying bitvector encoders. The inputs:

- shakes, scifi and bible: excerpts from Shakespeare's plays, a science-fiction novel, and the Bible, all in English;
- chr7, chr22, and coli: genome sequences from the human Chromosome 7 and 22, and E-coli bacteria, downloaded from the UCSC Genome Browser [40];
- euler, pi_1m, and pi_10m: first 2 million digits of the Euler constant, and 1 and 10 million digits of π.

The results are in Table II. It seems that *on real text inputs random access is consistently faster with R3D3 than with RRR at moderate block sizes* while rank performance is similar, and even at $b = 256$ we see only a minor performance hit. This is most probably due to the larger inputs and thereby CPU cache performance dominating query times. Interestingly, access ran slower than the more complex rank queries.

This experiment spectacularly highlights the benefits of data compression for operating on large quantities of data: it simultaneously delivers significant space savings over an uncompressed representation *and* implements fast operations on the content, permitting powerful queries of the type "How many times digit 5 occurs in π until the $500,000$-th position?" ($rank_5(\pi, 500000)$) or "Which is the 500-th valid ZIP code?" ($\mathrm{select}_1(\mathrm{zip}, 500)$), which a naive uncompressed representation does not even support out of the box.

These observations are further confirmed by our experiments on real Internet forwarding tables (see Table III). We used the *XBW* scheme of [31], a pair of a bitvector and a wavelet tree that together encode a prefix tree, to compress real FIB instances taken from operational Internet routers. Again, R3D3 approaches the entropy at larger block sizes and beats RRR multiple times, and it supports longest-prefix matches faster than the RRR-based encoding at roughly 5 times the speed as reported in [31].

V. CONCLUSION

Throughout the recent years, compressed data structures have gained wide-spread adoption in information retrieval, computational geometry, bioinformatics, networking, and big data. This is on the one hand due to their potential for making it possible to operate on unprecedentedly huge instances of data and, on the other hand, because they support much more complex queries to the compressed data, like rank and select, with zero performance impact. In many cases compression creates a win-win situation, as the memory footprint of large bodies of information can be freely decreased and meanwhile processing may even get faster, thanks to the data drifting closer to the CPU in the cache hierarchy.

In this paper, we have proposed R3D3 as a new tool for compressing and indexing bitvectors. R3D3 is, in contrast to previous work, doubly opportunistic, in that it realizes substantial space savings on the compressed data and the index alike. Furthermore, it allows to strike a fine space–time balance as required by the application at hand, with a smooth transition between the extremes. We have shown that most benefits already manifest themselves at moderate block sizes, realizing several times smaller encodings at only a slight performance impact compared to the state-of-the-art compressed bitvector scheme, RRR. At the extreme, for very large blocks R3D3 may provide 10-fold space reduction over uncompressed data and over RRR, in exchange of at most 50% performance penalty. Notably, on real data R3D3 proved faster than RRR. And because underlying most data indexing schemes, like compressed text indexes or compressed labeled trees, there is a bitvector data structure behind the scenes, the benefits of R3D3 also appear when compressing complex information, like small entropy textual data or genomes.

REFERENCES

[1] J. Manyika, M. Chui, B. Brown, J. Bughin, R. Dobbs, C. Roxburgh, and A. H. Byers, "Big data: The next frontier for innovation, competition, and productivity," McKinsey Global Institute, Tech. Rep., June 2011.

[2] O. Trelles, P. Prins, M. Snir, and R. C. Jansen, "Big data, but are we ready?" *Nat Rev Genet*, vol. 12, no. 3, p. 224, 2009.

[3] D. E. Knuth, *The Art of Computer Programming: Sorting and Searching*, ser. Series in Computer Science. Addison Wesley, 1998.

[4] R. Agarwal, A. Khandelwal, and I. Stoica, "Succinct: Enabling queries on compressed data," in *Proceedings of the 12th USENIX Conference on Networked Systems Design and Implementation*, ser. NSDI'15, 2015, pp. 337–350.

[5] G. Navarro and V. Mäkinen, "Compressed full-text indexes," *ACM Comput. Surv.*, vol. 39, no. 1, 2007.

[6] G. Jacobson, "Space-efficient static trees and graphs," in *Proceedings of the 30th Annual Symposium on Foundations of Computer Science*, ser. SFCS '89, 1989, pp. 549–554.

[7] W.-K. Hon, K. Sadakane, and W.-K. Sung, "Succinct data structures for searchable partial sums," in *Algorithms and Computation: 14th International Symposium*, 2003, pp. 505–516.

[8] C. K. Poon and W. K. Yiu, "Opportunistic data structures for range queries," in *Computing and Combinatorics: 11th Annual International Conference*, 2005, pp. 560–569.

[9] J. Barbay, M. He, J. I. Munro, and S. R. Satti, "Succinct indexes for strings, binary relations and multilabeled trees," *ACM Trans. Algorithms*, vol. 7, no. 4, pp. 1–27, 2011.

[10] D. Benoit, E. D. Demaine, J. I. Munro, R. Raman, V. Raman, and S. S. Rao, "Representing trees of higher degree," *Algorithmica*, vol. 43, no. 4, pp. 275–292, 2005.

[11] R. F. Geary, R. Raman, and V. Raman, "Succinct ordinal trees with level-ancestor queries," *ACM Trans. Algorithms*, vol. 2, no. 4, pp. 510–534, 2006.

[12] V. R. R. Raman and S. R. Satti, "Succinct indexable dictionaries with applications to encoding k-ary trees, prefix sums and multisets." *ACM Transactions on Algorithms*, vol. 3(4), 2007.

[13] P. Ferragina, F. Luccio, G. Manzini, and S. Muthukrishnan, "Compressing and indexing labeled trees, with applications," *J. ACM*, vol. 57, no. 1, pp. 1–33, 2009.

[14] F. Claude and G. Navarro, "A fast and compact web graph representation," in *String Processing and Information Retrieval: 14th International Symposium*, 2007, pp. 118–129.

[15] P. Elias, "Efficient storage and retrieval by content and address of static files," *J. Assoc. Comput. Mach.*, vol. 21, pp. 246–260, 1974.

[16] R. Grossi, A. Gupta, and J. S. Vitter, "High-order entropy-compressed text indexes," in *Proceedings of the Fourteenth Annual ACM-SIAM Symposium on Discrete Algorithms*, ser. SODA '03, 2003, pp. 841–850.

[17] A. Golynski, J. I. Munro, and S. S. Rao, "Rank/select operations on large alphabets: A tool for text indexing," in *Proceedings of the Seventeenth Annual ACM-SIAM Symposium on Discrete Algorithm*, ser. SODA '06, 2006, pp. 368–373.

[18] P. Ferragina, G. Manzini, V. Mäkinen, and G. Navarro, "Compressed representations of sequences and full-text indexes," *ACM Trans. Algorithms*, vol. 3, no. 2, 2007.

[19] T. Gagie, G. Navarro, and S. J. Puglisi, "New algorithms on wavelet trees and applications to information retrieval," *Theor. Comput. Sci.*, vol. 426-427, pp. 25–41, 2012.

[20] P. Ferragina and G. Manzini, "Opportunistic data structures with applications," in *Proceedings of the 41st Annual Symposium on Foundations of Computer Science*, ser. FOCS '00, 2000.

[21] K. Sadakane, "Succinct data structures for flexible text retrieval systems," *J. of Discrete Algorithms*, vol. 5, no. 1, pp. 12–22, 2007.

[22] S. Gog, "Compact and succinct data structures: From theory to practice," available online: http://es.csiro.au/ir-and-friends/20131111/anu_gog_seminar.pdf, 2015.

[23] M. C. Brandon, D. C. Wallace, and P. Baldi, "Data structures and compression algorithms for genomic sequence data," *Bioinformatics*, vol. 25, no. 14, pp. 1731–1738, 2009.

[24] S. Kuruppu, B. Beresford-Smith, T. Conway, and J. Zobel, "Iterative dictionary construction for compression of large DNA data sets," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 9, no. 1, pp. 137–149, Jan 2012.

[25] R. Raman, "Succinct data structures for data mining," Workshop on Algorithms for Large-Scale Information Processing in Knowledge Discovery, 2014.

[26] P. B. Miltersen, "Lower bounds on the size of selection and rank indexes," in *Proceedings of the Sixteenth Annual ACM-SIAM Symposium on Discrete Algorithms*, ser. SODA '05, 2005, pp. 11–12.

[27] D. Okanohara and K. Sadakane, "Practical entropy-compressed rank/select dictionary," in *Proceedings of the Meeting on Algorithm Engineering & Expermiments*, 2007, pp. 60–70.

[28] S. Vigna, "Broadword implementation of rank/select queries," in *Proceedings of the 7th International Conference on Experimental Algorithms*, ser. WEA'08, 2008, pp. 154–168.

[29] F. Claude and G. Navarro, "Practical rank/select queries over arbitrary sequences," in *Proceedings of the 15th International Symposium on String Processing and Information Retrieval*, ser. SPIRE '08, 2009, pp. 176–187.

[30] G. Navarro and E. Providel, "Fast, small, simple rank/select on bitmaps," in *Experimental Algorithms: 11th International Symposium*, 2012, pp. 295–306.

[31] G. Rétvári, J. Tapolcai, A. Kőrösi, A. Majdán, and Z. Heszberger, "Compressing IP forwarding tables: towards entropy bounds and beyond," in *ACM SIGCOMM 2013*, 2013, pp. 111–122.

[32] T. M. Cover and J. A. Thomas, *Elements of information theory*. Wiley-Interscience, 1991.

[33] S. Gog, "Succinct Data Structure Library," https://github.com/simongog/sdsl-lite.

[34] D. E. Knuth, *The Art of Computer Programming: Combinatorial Algorithms*, ser. Series in Computer Science. Addison-Wesley, 2011.

[35] V. Makinen and G. Navarro, "Dynamic entropy-compressed sequences and full-text indexes," *ACM Transactions on Algorithms*, vol. 4, 2008.

[36] A. Majdán and G. Rétvári, "Development and performance evaluation of fast combinatorial unranking implementations," in *EUNICE*, 2014.

[37] S. Vigna, "Quasi-succinct indices," in *ACM International Conference on Web Search and Data Mining, WSDM*, 2013, pp. 83–92.

[38] M. Nagy, "Github homepage," https://nmate.github.io, 2019.

[39] I. Witten, T. Bell, and J. Cleary, "The Calgary Corpus," 1987, http://corpus.canterbury.ac.nz/descriptions/#calgary.

[40] UCSC Genome Bioinformatics, "The UCSC Genome Browser," http://hgdownload.cse.ucsc.edu/downloads.html.

APPENDIX

Proof of Lemma 1: The following metadata are stored in the RRR index for each superblock i and each block j:

- P_i: the address of the ith superblock;
- R_i: cummulative rank up to the ith superblock;
- L_{ij}: relative address of block j inside superblock i;
- Q_{ij}: relative rank of block j inside superblock i;
- K_{ij}: the block class $c_j = \mathrm{popcount}(b_j)$.

Both P and R require $\frac{n}{s}\log(n)$ bits, K needs $\frac{n}{b}\log(b)$ bits, while L and Q, both holding values relative to the containing superblock, can use $\log(s)$ bits per block. In total

$$M_I = 2\frac{n}{s}\log n + \frac{n}{b}\log b + 2\frac{n}{b}\log s \quad (2)$$

$$= 2\frac{n}{b}\left(\frac{b}{s}\log n + \frac{\log b}{2} + \log s\right) \quad . \quad (3)$$

Now, (2) gives a useful hint on how to select the superblock size: introducing the notation $x = \frac{b}{s}$ we see that (2) is minimal where $\frac{\partial}{\partial x}\frac{n}{b}\left(x\log n + \frac{\log b}{2} + \log\frac{b}{x}\right) = 0$, which gives $x = \frac{1}{\log n}$ and hence for the superblock size $s = b\log n$. With this setting, we get $M_I = n/b(2+3\log(b)+2\log\log n)$ as required by the claim of the Lemma. ■

Proof of Lemma 2: First, we observe that instead of calculating the space occupancy of each block b_i one by one, it is enough to deal with the size of an "average" block with $c = pb$ (the proof is trivial using Jensen's inequality, we omit the details). The UBA stores a bit for each bucket plus another bit for each bit set in the block, yielding $2^{\log c} + c = 2^{\log b - l} + c = \frac{b}{2^l} + c$ bits overall, while the LBA consists of c elements, each of l bits. Summed up for each of the $\frac{n}{b}$ blocks:

$$M_D = \frac{n}{b}\left(\frac{b}{2^l} + c + lc\right) = n\left(2^{-l} + p + pl\right) \quad . \quad (4)$$

Recall that the choice for parameter l is elemental in EF; usually $l = \lfloor\log\frac{b}{c}\rfloor$. First, to demonstrate the main idea of the proof we give the treatment for the simplified case when we omit rounding to integers, then we discuss how to handle this discrepancy. Letting $l = \log b - \log c$ firstly yields

$$M_D = n\left(\frac{c}{b} + p + p\log\frac{b}{c}\right) = n\left(p + p + p\log\frac{1}{p}\right) \quad (5)$$

$$\leq n\left(p + (1-p)\log\frac{1}{1-p} + p\log\frac{1}{p}\right) = np + nH_0 \quad , \quad (6)$$

by that $p \leq (1-p)\log(\frac{1}{1-p})$ if $p \in (0, \frac{1}{2}]$, as requested.

Secondly, taking care of integrality $l = \lfloor\log\frac{b}{c}\rfloor$ and using that $\frac{b}{c} = \frac{1}{p}$, we write:

$$M_D = n\left(2^{-\lfloor\log\frac{1}{p}\rfloor} + p + p\left\lfloor\log\frac{1}{p}\right\rfloor\right) \quad . \quad (7)$$

To prove the statement, it is now enough to show that (5) is larger than, or equal to (7), or, equivalently, that the difference

$$2^{-x} + xp - (2^{-\lfloor x\rfloor} + \lfloor x\rfloor p)$$

is non-negative, where we used the shorthand $x = \log(\frac{1}{p})$. Clearly for $0 \leq x < 1$ the difference equals $2^{-x} + xp - 1$, which is always positive as $2^{-x} \geq 1$ in this range and x and p are positive. Next, we will show that $f(x) = 2^{-x} + xp$ is a decreasing function of x for $x \geq 1$. Substitute $p = 2^{-x}$ to get

$$f(x) = 2^{-x} + x2^{-x} = 2^{-x}(1+x)$$

Finally, we need to show that the derivate is negative:

$$\frac{\partial f(x)}{\partial x} = 2^{-x} - 2^{-x}(1+x)\ln(2) = 2^{-x}\big(1 - (1+x)\ln(2)\big).$$

Clearly, 2^{-x} is positive and $1-(1+x)\ln(2)$ is negative for $x > \frac{1}{\ln(2)} - 1 \approxeq 0.44$. This completes the proof. ■

Proof of Theorem 2: We only need to show that $M_I + M_D$ is as required. Write $M_I = \frac{2n}{b} + \frac{3n}{b}\log b + \frac{2n}{b}\log\log n$ and substitute $b = \frac{C}{p}$ to get $\frac{2np}{C} + \frac{3np}{C}\log\frac{C}{p} + \frac{2pn}{C}\log\log n$. Using that $p \leq \frac{1}{2}$ and so $p\log\frac{1}{p} \leq H_0$ and $2p \leq H_0$, we write for the first component $\frac{2pn}{C} \leq nH_0\frac{1}{C}$, for the second $\frac{3np}{C}\log\frac{C}{p} = \frac{n}{C}\left(3p\log\frac{1}{p} + 3p\log C\right) \leq \frac{n}{C}3H_0 + \frac{n}{C}3p\log C = nH_0O(\frac{\log C}{C})$, and for the third $\frac{n}{C}2p\log\log n = \frac{n}{C}2pO(\log C)$ (by that $C = O(\log n)$) and thus $nH_0O(\frac{\log C}{C})$ using the same substitutions as before. Thus, $M_I = nH_0O(\frac{\log C}{C})$ and $M_D = nH_0 + np \leq nH_0 + \frac{1}{2}nH_0$, yielding the overall size $M_I + M_D = nH_0 + nH_0\left(\frac{1}{2} + O(\frac{\log C}{C})\right)$ bits, which completes the proof. ■

Máté Nagy received the M.Sc. degree in electrical engineering from the Budapest University of Technology and Economics in 2010, where he is currently pursuing the Ph.D. degree with the High Speed Network Laboratory (HSN Lab). He has participated in several research projects driven by Ericsson Research, Hungary. His research interests span routing, software defined networking and information theory.

János Tapolcai received the M.Sc. degree in technical informatics and the Ph.D. degree in computer science from the Budapest University of Technology and Economics (BME), Budapest, in 2000 and 2005, respectively, and the D.Sc. degree in engineering science from the Hungarian Academy of Sciences (MTA) in 2013. He is currently a Full Professor with the High-Speed Networks Laboratory, Department of Telecommunications and Media Informatics, BME. He has authored over 150 scientific publications. His current research interests include applied mathematics, combinatorial optimization, optical networks and IP routing, addressing, and survivability. He was a recipient of several Best Paper Awards, including ICC'06, DRCN'11, HPSR'15, and NaNa'16. He is a winner of the MTA Lendület Program and the Google Faculty Award in 2012, Microsoft Azure Research Award in 2018. He is a TPC member of leading conferences, e.g. IEEE INFOCOM 2012-2017, and the general chair of ACM SIGCOMM 2018.

Gábor Rétvári received the M.Sc. and Ph.D. degrees in electrical engineering from the Budapest University of Technology and Economics in 1999 and 2007. He is now a Senior Research Fellow at the Department of Telecommunications and Media Informatics. His research interests include all aspects of network routing and switching, the programmable data plane, and the networking applications of computational geometry and information theory. He maintains several open source scientific tools written in Perl, C, and Haskell.

IEEE ICC®

IEEE INTERNATIONAL CONFERENCE ON COMMUNICATIONS
7 - 11 June 2020 // Dublin, Ireland
Communications Enabling Shared Understanding

IEEE ComSoc IEEE Communications Society

CALL FOR PAPERS

The 2020 IEEE International Conference on Communications (ICC 2020) will be held in the exciting city of Dublin, Ireland, from 7-11 June 2020. Themed, "Communications Enabling Shared Understanding," this flagship conference of the IEEE Communications society will feature a comprehensive high-quality technical program including 13 symposia and a variety of tutorials and workshops. IEEE ICC 2020 will also include an attractive Industry program aimed at practitioners, with keynotes and panels from prominent research, industry and government leaders, business and industry panels, and vendor exhibits.

TECHNICAL SYMPOSIA

- Ad Hoc and Sensor Networks
- Cognitive Radio and AI-Enabled Networks
- Communication and Information Systems Security
- Communications QoS, Reliability and Modeling
- Communications Software, Services & Multimedia Apps.
- Communication Theory
- Green Communication Systems and Networks
- Next Generation Networking and Internet
- Optical Networks and Systems
- Signal Processing for Communications
- Wireless Communications
- Mobile and Wireless Networks
- Selected Areas in Communications
  - Access Networks/Systems and Power Line Communications
  - Big Data
  - Cloud & Fog Computing, Networking, and Storage
  - Smart Grid Communications
  - E-Health
  - Internet of Things
  - Molecular, Biological, and Multi-Scale Communications
  - Satellite and Space Communications
  - Social Networks
  - Tactile Internet

INDUSTRY FORUMS AND EXHIBITION PROGRAM

IEEE ICC 2020 will feature several prominent keynote speakers, major business and technology forums, and a large number of vendor exhibits.

TUTORIALS

Proposals are invited for half- or full-day tutorials in all communication and networking topics.

WORKSHOPS

Proposals are invited for half- or full-day workshops in all communication and networking topics.

IMPORTANT DATES

Paper Submission: 14 October 2019
Acceptance Notification: 27 January 2020
Camera-Ready: 24 February 2020
Tutorial Proposals: 7 October 2019
Workshop Proposals: 5 August 2019

Full details of submission procedures are available at **icc2020.ieee-icc.org**

GENERAL CHAIR
Declan Ganley (Rivada Networks, Ireland)

GENERAL VICE-CHAIR
Linda Doyle (Trinity College Dublin, Ireland)

EXECUTIVE CHAIR
Brendan Jennings (Waterford Institute of Technology, Ireland)

EXECUTIVE VICE-CHAIR
Dirk Pesch (University College Cork, Ireland)

INDUSTRY PROGRAM CO-CHAIRS
Joe Betser (Aerospace Corporation, USA)
Sven van der Meer (Ericsson, Ireland)

PATRONAGE CHAIR
Alan Davy (Waterford Institute of Technology, Ireland)

FINANCE CHAIR
Martin Johnsson (Waterford Institute of Technology, Ireland)

TECHNICAL PROGRAM CHAIR
Luiz DaSilva (Trinity College Dublin, Ireland)

TECHNICAL PROGRAM CO-CHAIRS
Jianwei Wang (Chinese Univ. of Hong Kong)
Min Song (Stevens Institute of Technology, USA)

TUTORIALS CO-CHAIRS
Wenye Wang (NC State University, USA)
Nicola Marchetti (Trinity College, Ireland)

WORKSHOPS CO-CHAIRS
Zhisheng Niu (Tsinghua University, China)
Allen MacKenzie (Virginia Tech, USA)
Sofie Pollin (KU Leuven, Belgium)

PUBLICATIONS CO-CHAIRS
Angela Zhang (Chinese Univ. of Hong Kong)
Johann Marquez-Barja (U. Antwerp, Belgium)

STUDENT TRAVEL GRANTS CO-CHAIRS
Rose Hu (Utah State University, USA)
Michele Nogueira (UFPR, Brazil)

Guidelines for our Authors

Format of the manuscripts

Original manuscripts and final versions of papers should be submitted in IEEE format according to the formatting instructions available on

https://journals.ieeeauthorcenter.ieee.org/
Then click: "IEEE Author Tools for Journals"
- "Article Templates"
- "Templates for Transactions".

Length of the manuscripts

The length of papers in the aforementioned format should be 6-8 journal pages.
Wherever appropriate, include 1-2 figures or tables per journal page.

Paper structure

Papers should follow the standard structure, consisting of *Introduction* (the part of paper numbered by "1"), and *Conclusion* (the last numbered part) and several *Sections* in between.
The Introduction should introduce the topic, tell why the subject of the paper is important, summarize the state of the art with references to existing works and underline the main innovative results of the paper. The Introduction should conclude with outlining the structure of the paper.

Accompanying parts

Papers should be accompanied by an *Abstract* and a few *index terms (Keywords)*. For the final version of accepted papers, please send the short cvs and *photos* of the authors as well.

Authors

In the title of the paper, authors are listed in the order given in the submitted manuscript. Their full affiliations and e-mail addresses will be given in a footnote on the first page as shown in the template. No degrees or other titles of the authors are given. Memberships of IEEE, HTE and other professional societies will be indicated so please supply this information. When submitting the manuscript, one of the authors should be indicated as corresponding author providing his/her postal address, fax number and telephone number for eventual correspondence and communication with the Editorial Board.

References

References should be listed at the end of the paper in the IEEE format, see below:

a) Last name of author or authors and first name or initials, or name of organization
b) Title of article in quotation marks
c) Title of periodical in full and set in italics
d) Volume, number, and, if available, part
e) First and last pages of article
f) Date of issue
g) Document Object Identifier (DOI)

[11] Boggs, S.A. and Fujimoto, N., "Techniques and instrumentation for measurement of transients in gas-insulated switchgear," IEEE Transactions on Electrical Installation, vol. ET-19, no. 2, pp.87–92, April 1984. DOI: 10.1109/TEI.1984.298778

Format of a book reference:

[26] Peck, R.B., Hanson, W.E., and Thornburn, T.H., Foundation Engineering, 2nd ed. New York: McGraw-Hill, 1972, pp.230–292.

All references should be referred by the corresponding numbers in the text.

Figures

Figures should be black-and-white, clear, and drawn by the authors. Do not use figures or pictures downloaded from the Internet. Figures and pictures should be submitted also as separate files. Captions are obligatory. Within the text, references should be made by figure numbers, e.g. "see Fig. 2."
When using figures from other printed materials, exact references and note on copyright should be included. Obtaining the copyright is the responsibility of authors.

Contact address

Authors are requested to submit their papers electronically via the EasyChair system. The link for submission can be found on the journal's website:

www.infocommunications.hu/for-our-authors

If you have any question about the journal or the submission process, please do not hesitate to contact us via e-mail:

Pál Varga – Editor-in-Chief:
pvarga@tmit.bme.hu

Rolland Vida – Associate Editor-in-Chief:
vida@tmit.bme.hu

Special Issue

of the **Infocommunication Journal**

Call for Papers

"Cognitive Infocommunications Theory and Applications"

Cognitive infocommunications (CogInfoCom) investigates the link between the research areas of infocommunications and cognitive sciences, as well as the various engineering applications which have emerged as the synergic combination of these sciences. The primary goal of CogInfoCom is to provide a systematic view of how cognitive processes can co-evolve with infocommunications devices so that the capabilities of the human brain may not only be extended through these devices, irrespective of geographical distance but may also be blended with the capabilities of any artificially cognitive system. This merging and extension of cognitive capabilities are targeted towards engineering applications in which artificial and/or natural cognitive systems are enabled to work together more effectively.

This special issue collects the latest results emerging on the field of Cognitive Infocommunications.

Guest Editor:

Prof. Peter Baranyi DSc.,
Budapest University of Technology and Economics

Important dates:

Submission deadline: **1st of September, 2019**
Notification of the first review: **1st of November, 2019**
Deadline for revision is **1st of December 2019**
Camera Ready submission is **10 of January, 2020**

Regarding manuscript submission information, please visit:
https://www.infocommunications.hu/for-our-authors

SCIENTIFIC ASSOCIATION FOR INFOCOMMUNICATIONS

Who we are

Founded in 1949, the Scientific Association for Infocommunications (formerly known as Scientific Society for Telecommunications) is a voluntary and autonomous professional society of engineers and economists, researchers and businessmen, managers and educational, regulatory and other professionals working in the fields of telecommunications, broadcasting, electronics, information and media technologies in Hungary.

Besides its 1000 individual members, the Scientific Association for Infocommunications (in Hungarian: Híradástechnikai Tudományos Egyesület, HTE) has more than 60 corporate members as well. Among them there are large companies and small-and-medium enterprises with industrial, trade, service-providing, research and development activities, as well as educational institutions and research centers.

HTE is a Sister Society of the Institute of Electrical and Electronics Engineers, Inc. (IEEE) and the IEEE Communications Society.

What we do

HTE has a broad range of activities that aim to promote the convergence of information and communication technologies and the deployment of synergic applications and services, to broaden the knowledge and skills of our members, to facilitate the exchange of ideas and experiences, as well as to integrate and harmonize the professional opinions and standpoints derived from various group interests and market dynamics.

To achieve these goals, we…

- contribute to the analysis of technical, economic, and social questions related to our field of competence, and forward the synthesized opinion of our experts to scientific, legislative, industrial and educational organizations and institutions;
- follow the national and international trends and results related to our field of competence, foster the professional and business relations between foreign and Hungarian companies and institutes;
- organize an extensive range of lectures, seminars, debates, conferences, exhibitions, company presentations, and club events in order to transfer and deploy scientific, technical and economic knowledge and skills;
- promote professional secondary and higher education and take active part in the development of professional education, teaching and training;
- establish and maintain relations with other domestic and foreign fellow associations, IEEE sister societies;
- award prizes for outstanding scientific, educational, managerial, commercial and/or societal activities and achievements in the fields of infocommunication.

Contact information

President: **GÁBOR MAGYAR, PhD** • *elnok@hte.hu*

Secretary-General: **ERZSÉBET BÁNKUTI** • *bankutie@ahrt.hu*

Operations Director: **PÉTER NAGY** • *nagy.peter@hte.hu*

International Affairs: **ROLLAND VIDA, PhD** • *vida@tmit.bme.hu*

Address: H-1051 Budapest, Bajcsy-Zsilinszky str. 12, HUNGARY, Room: 502

Phone: +36 1 353 1027

E-mail: *info@hte.hu*, Web: *www.hte.hu*