Recent Trends in Communication Networks

Edited by Pinaki Mitra

Recent Trends in Communication Networks
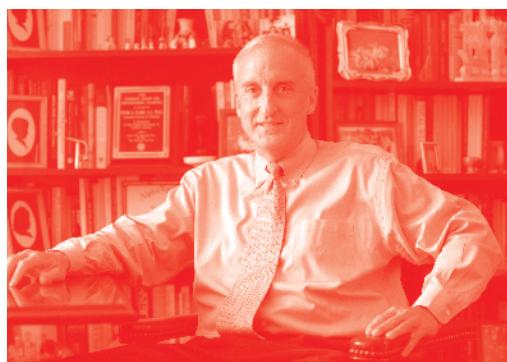
Edited by Pinaki Mitra

IntechOpen

Supporting open minds since 2005

We are IntechOpen,
the world's leading publisher of Open Access books
Built by scientists, for scientists

4,900+
Open access books available

124,000+
International authors and editors

140M+
Downloads

151
Countries delivered to

Our authors are among the
Top 1%
most cited scientists

12.2%
Contributors from top 500 universities

BOOK CITATION INDEX
CLARIVATE ANALYTICS
INDEXED

WEB OF SCIENCE™

Selection of our books indexed in the Book Citation Index
in Web of Science™ Core Collection (BKCI)

Interested in publishing with us?
Contact book.department@intechopen.com

Numbers displayed above are based on latest data collected.
For more information visit www.intechopen.com

Meet the editor

Pinaki Mitra is currently an associate professor at the Department of Computer Science and Engineering, IIT Guwahati. He obtained his B. Tech in Computer Science and Engineering from Jadavpur University, Kolkata in 1987, India; and his M. Tech in Computer Science and Engineering from the Indian Institute of Science Bangalore, India in 1989. He obtained his Ph.D. from the Simon Fraser University, Canada in 1994. He has worked on a project at the Department of Computer Science and Engineering, Jadavpur University. Subsequently, he joined the National Institute of Management, Kolkata, and served as an assistant professor. He joined IIT Guwahati in December, 2004. His research interests include cryptography, network security, computer graphics, and multimedia.

Contents

Preface

The book focuses on different areas of recent developments in communication research. Topics that are covered in the book comprise secure communication, error control, different generations of communication networks, software-defined networks, communication protocols, circuit design issues for communication, and multimedia data communication. All these research methodologies indicate how computing and communication technology complement each other. Communication technology has greatly enhanced computing power from stand-alone computers to networked systems to distributed systems. Advanced computing techniques have supplemented this development, thus ensuring secure, error-free, and uninterrupted transmission in different communication environments. The book presents different recent research problems in the area of communication. The content of the book will be beneficial for future generations of communication research.

I would like to express my sincere thanks to my postdoctoral fellow Dr. Behrouz Zolfaghari for carrying out several reviews that helped me in editing the book. Last but not least I would like to convey my sincere thanks to Mrs. Dolores Kuzelj, the Author Service Manager, who had provided constant support in the entire editing process of the book.

Pinaki Mitra
Professor,
Department of Computer Science and Engineering,
Guwahati, India

Section 1

Introductory Chapter

Introductory Chapter: Recent Trends in Communication Networks

Pinaki Mitra

1. Introduction

With the emergence of distributed computing platforms and computer networks in the last few decades, there is almost no stand-alone computer anymore. In the last few decades, stand-alone computers in a distributed system communicate in order to share computing power. Communication in computer networks was initially aiming at resource sharing, but nowadays it serves to a variety of functions including computing power sharing. There has been plenty of research works in the area of communications focusing on the development of new networking protocols, transmission media, and different generations of communication technology.

Figure 1 shows the communication between two hosts in its simplest form. Hosts in computer networks are no longer confined to computers; they can be cell phones, tablets, smart sensors, and so on. In **Figure 1**, the function of the modem is to perform modulation and demodulation.

During the course of time, different types of networks evolved. The initially existing telephone network was used for the purpose of computer communication. This was circuit-switched. There was no buffering in intermediate nodes. Subsequently message-switched network evolved that used the concept of storing in buffer and forwarding. Then packet-switched network came, where messages are fragmented into small-sized packets that can be stored in primary memory, unlike earlier messages that needed secondary storage. Subsequently, we had frame relay, ATM, and virtual circuit packet switching that is a combination of circuit switching and packet switching.

Moreover, a variety of multiplexing techniques are used in data communications. Some commonly used multiplexing techniques are frequency division multiplexing or FDM, time division multiplexing or TDM, and statistical time-division multiplexing. Frequency-division multiplexing used modulation at sender end and demodulation at the receiving end. But it was not very popular like TDM. But in the recent generation of communication, networks are using orthogonal

Figure 1.
Illustration of data communication in computer networks.

frequency-division multiplexing (OFDM). Also in communication technology, we had the development of GSM, spread spectrum, and CDMA.

Another aspect of data communication which has witnessed a variety of developments is the technology of the transmission media. It can be either wired or wireless. For wired transmission media, we had twisted pair, coaxial, and fiber-optic cables. Research had been carried out to reduce signal distortion due to attenuation, delay, and interference of different types of noise. Equalization techniques are used to avoid different distortions. In wireless communication, we use microwave communication or satellite communication using geostationary satellites.

The next important issue related to the communication network is the development of network protocols. Initially, we had seven-layered ISO/OSI protocols. After that five-layered TCP/IP protocol evolved. The purpose of these protocols was to ensure error control, flow control, and routing and congestion control. In these protocols, there is a virtual communication between corresponding layers in two different nodes. This takes place through the appending of packet headers, while the information moves down the layer at the source node and removal/stripping of the packet header by the corresponding layer, while the information flows up in the destination node. With recent developments of the Internet of things (IOT), a modified version of TCP/IP protocol evolved named 6LoWPAN. These protocols are not only implemented in the communicating nodes/computers of the communication network but also at routers and switches.

With the increased use of communication networks in day to day life, they became susceptible to attacks from malicious users. There are usually two different types of attacks, namely, passive attacks and active attacks. In passive attacks, the transmitted information is leaked to the eavesdropper but the transmitted message remains unaltered. This can be avoided by using sophisticated encryption and decryption methodologies. In contrast, active attacks involve modification of message content, masquerade, non-repudiation, and denial of service. The modification of message content can be handled using different error detection and error correction methods. There had been several works related to different types of coding in the field of communication research. Non-repudiation can be handled by the cryptographic technique of digital signatures or message authentication code (MAC). Denial of service involves flooding of networks with garbage packets from malicious users. Several machine learning techniques for outlier or anomaly detection are used to detect such attacks.

Performance evaluation of different communication techniques can be carried out using the following four quality of service or QoS parameters. They are throughput, packet loss, delay, and jitter/delay variation. We have to suitably adjust the trade-off between these parameters to optimize the quality of service.

Nowadays, we use several mini handheld devices that use the features of the mobile communication network. These devices usually have very limited storage, computing power, and energy. For this purpose, several types of research had been carried out to reduce the computation and communication overheads without compromising accuracy, security, etc. Another important difference in the area of the mobile communication networks is unlike earlier networks whose topology is always fixed; here we have to handle the problem of variable topology. The interconnecting link between two nodes of the network may or may not exist during the course of time. Usually, each node in the network has a limited transmission range. Connections are established between two mobile nodes that are within the range. These links will cease to exist when these two nodes go out of range of each other. These networks are also called MANET or mobile ad hoc networks. There are several issues that are different for routing of packets in MANET as compared to fixed interconnection networks. Several types of research had been carried out

for the design of routing protocols in MANET that are also susceptible to various attacks. A recent development in the area of MANET is the vehicular ad hoc networks (VANET). Since a huge amount of data and computation is involved here, several techniques from big data analytics are employed to improve efficiency.

Recently the development of multimedia had enhanced developments of various existing disciplines of computing like operating systems, database, and communication. Multimedia data involves a mixture of image, text, audio, video, and graphics data. These are usually referred to as multimedia content. In the field of operating systems, the advent of multimedia had led to the development of real-time operating systems. Different scheduling algorithms like the earliest deadline first or EDF and rate monotonic algorithm or RMA are adopted for these applications. In the field of the database due to multimedia, we had seen the development of object-relational DBMS where we can efficiently perform content-based retrieval. In the area of communication, the advent of multimedia had led to video on demand. The communication protocol that is used for multimedia communication is RTP that usually runs over UDP that is a simplified version of TCP. We had seen the development of different types of a content distribution network or CDN for this purpose. Since in communication video data involve large uplink/downlink bandwidth, research had been carried out to transmit these data as a mix of different definitions. High-definition video is meant for nodes having large bandwidth for downlink purposes. Medium-definition video is meant for nodes having moderate bandwidth for downlink purposes. Low-definition video is meant for nodes having low bandwidth for downlink purposes. In content distribution network in addition to four quality of service, an additional parameter is introduced that measures the user's feeling or experience of the displayed video satisfying the bandwidth constraint. This is also known as the Quality of Experience (QoE). Usually, this parameter is quantified using peak signal-to-noise ratio or PSNR value of the displayed video stream.

Recent developments in different branches of communications technologies have led to hybrid technologies such as mobile satellite communication (SATCOMM) which are finding their applications in several environments such as unmanned aerial vehicle (UAV) systems.

In recent years, there has been a fruitful interaction between computer network technology and other emerging technologies. Computer networks are now used to provide cloud services. They are interacting with fog computing technology to build content delivery networks. The latter networks provide a lot of services including social networking, digital cinema, etc. But with the development of the ecosystem of these networks, a variety of challenges and requirements are raised among which one can refer to security and specially copyright protection issues.

For recent developments related to communication and networks readers are suggested to refer to [1–4]. For recent developments in the area of network security, readers are suggested to refer to [5].

2. Conclusion

In the field of computers with the advent of the Internet, the topic of computer communication gained significant importance. Different modulation and demodulation techniques evolved for improved long-distance communication. Switching techniques handled efficient data transmission. Different generations of communication networks like 3G, 4G, 5G, and so on evolved in the course of time. Different communication network protocols evolved. These networks got susceptible to various types of attacks. The theory of cryptography and coding theory evolved to handle many

such problems. More recently with the advancement of mobile technologies and IoT, these algorithms had to take into consideration limited resources like battery power, storage, and processor capabilities. This had led to the development of new communication technologies for resource-constrained devices. Also in MANET due to time-varying interconnection, there is a necessity to maintain proper integrity and consistency in the distributed data. Due to large volume of data efficient computational techniques from big data analytics are adopted. Also in the recent development of multimedia communication, different adaptive streaming of multimedia content over different types of networks had evolved that would optimize different quality of service and quality of experience parameters. The book addresses these issues that arise in present-day research in the field of communication and networking and also presents several possible directions for future research.

Author details

Pinaki Mitra
Department of Computer Science and Engineering, IIT Guwahati, Guwahati, India

*Address all correspondence to: pinaki@iitg.ac.in

IntechOpen

References

[1] Stallings W. Data and Computer Communications. 8th ed. Pearson; 2006. ISBN: 0-13-243310-9

[2] Gupta P. Data Communications and Network. India: Prentice Hall; 2014. ISBN: 13: 978-8120348646

[3] Pal A. Data Communications and Computer Networks. India: Prentice Hall; 2014. ISBN: 9788120348455

[4] Bertsekas D, Gallager R. Data Networks. 2nd ed. India: Prentice Hall; 1992. ISBN: 0-13-200916-1

[5] Stallings W. Cryptography and Network Security: Principles and Practice. 4th ed. India: Prentice Hall; 2005. ISBN: 978-0-13-111502-6

Section 2

Security Issues in Communication Technology

Chapter 2

A Survey on Piracy Protection Techniques in Digital Cinema Watermarking Schemes

Behrouz Zolfaghari and Pinaki Mitra

Abstract

Watermarking is used in several areas such as CDNs (Content Delivery Networks), as part of the rights management system for counterfeit prevention. Watermarking schemes need some additional features in order to be used in digital cinema. In fact, extra watermarks are added to movies by cinema projectors in projection time, which help identify the cinema hall in which the illegal copy has been recorded. But distortions caused by hand vibrations and the point of view angle make it difficult to recover the watermark. This makes it necessary to be distortion-resistant for the watermarking schemes used in digital cinema. On the other hand, theatre owners would like to locate the camcorder that has recorded the pirate copy. This requires watermarking schemes to be able to estimate the distance and angle using the distributed pirate copy. In this chapter, we present a review on watermarking techniques specifically designed to attack the aforementioned problems.

Keywords: watermarking, piracy, digital cinema, copyright, survey, digital distribution

1. Introduction

Digital distribution has been considered by researchers since four decades ago [1–3]. With the growth of available bandwidth in the telecommunications infrastructures, it gradually appeared as one of the main purposes of communication during two decades. This encouraged a lot of research on different aspects of digital distribution over internet [4–6]. Digital distribution brings about several challenges among which one can refer to performance, but in this chapter, we are more interested in security. Security of digital distribution has been investigated from different aspects [7–9]. One of the most challenging aspects in digital distribution security is digital rights management, which has been the topic for several research works [10, 11]. With the emergence of internet, online digital distribution found its applications in a variety of environments [12–14]. Meanwhile, the public access to internet made security more and more challenging [15].

Content delivery is a branch of digital distribution, which is growing very fast in different aspects [16, 17]. The more content delivery systems grow, the more complex security their security management becomes [18]. Among advances in content delivery, one may mention CDNs [19–21]. These networks serve to high-performance delivery of digital contents as well as streaming of audio and video in collaboration with other technologies such as cloud computing [22, 23].

Different aspects of security in CDNs such as privacy [24], digital property management [25], intrusion detection [26], trust [27], and eavesdropping protection [28] have been studied by researchers.

Digital cinema is a service that can be provided using CDNs. It is an online video streaming and projection service gaining a research focus in recent years [29–31]. Security is a serious challenge in digital cinema [32–34] like other services provided over CDNs. One of the most important aspects of digital cinema security is digital rights management [35, 36], which is about preserving the publication rights for the generator of contents. Counterfeit protection is one of the most challenging problems in the area of digital cinema rights management.

Digital property management and especially counterfeit protection mostly depend on watermarking in a variety of environments [37–39]. Digital watermarking has been finding its applications in digital cinema in recent years [40, 41]. But, there is a delicate challenge faced by watermarking systems used in digital cinema. In the absence of proper protective measures, counterfeit copies of movies can be recorded by handheld cameras in the cinema hall. Movie theater holders are obliged to take preventive measures in this regard. This creates two problems, which need to be taken into consideration while designing watermarking schemes for digital cinema. These problems are explained in the following.

The first problem is faced by movie providers. When a pirate copy is distributed, providers would like to identify the theatre where the copy has been recorded. Then they can sue the identified theatres as they have not fulfilled their legal obligations. One way to approach this problem is to add watermarks to the movie by the projector in projection time. This watermark carries some identification information about the cinema hall as well as the projection time. This makes it possible to identify the hall in which a counterfeit copy has been recorded. But, the distortions made by hand vibrations and point of view angle, makes it difficult to recover the watermark from the movie. Thus, watermarking schemes designed for digital cinema should be specifically resistant against these distortions.

The second problem is to locate the seat on which the camcorder has been recording. This problem is to be solved by cinema holders as they are legally responsible for illegally recorded copies. Solving this problem requires digital cinema watermarking schemes to be able to calculate distance and angle.

There are some surveys on techniques used to solve the two problems mentioned above. But, they are outdated for such a fast-growing area [42, 43]. This motivates our work in this chapter, which is a comparative review on watermarking schemes used in digital cinema and the methods used by these schemes in order to achieve the mentioned goals.

The rest of this chapter is organized as follows. Section 2 presents an introduction to digital cinema. Section 3 discusses watermarking and related concepts. Section 4 presents a survey on watermarking in digital cinema. Section 5 concludes the chapter and suggests further research.

2. Digital cinema

Digital cinema refers to the set of processes, tools, and components which aim at the preparation, transmission, and projection of digital movies in cinema. This system is taking the place of traditional cinema which depends on reels of film being projected by film projectors.

In digital cinema, 2K (2048 × 1080 or 2.2 megapixels) or 4K (4096 × 2160 or 8.8 megapixels) resolutions are commonly used.
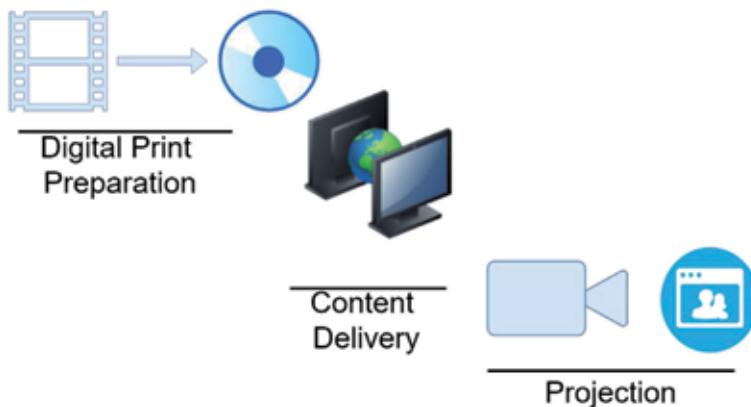
Figure 1.
A typical digital cinema system.

Among the advantages of digital cinema, we can point out reduction in time to market, reduction in cost per print, greater reach, and piracy control.

A digital cinema system consists of three main processes. In the first process, the film is converted to a digital print called DCP (Digital Cinema Package). The DCP is then delivered to the theatres and ingested (stored) by the related servers. This can be done through physical transportation of hard disks or through a digital distribution system. A content delivery network can serve to this purpose [44]. The DCP is often encrypted, and the keys required to decrypt the file are ingested in the screening tools to prevent theater owners from illegal screen extensions. In the third process, the movie is played back using a digital video projector. The aforementioned three processes are shown in **Figure 1**.

The first system specification for digital cinema was published in 2005 by a consortium of six major studios. This specification is referred to as DCI (Digital Cinema Initiatives) and uses JPG 2000 video encoding with XML playback standard.

3. Digital watermarking

Watermarking is a copyright ownership identification method used for noise-tolerant media such as video, image, and audio. In this technique, an intentional noise is added to the signal in the form of covertly embedded information. In most cases, the hidden information is normally not perceptible. Rather it requires some revealing process to be extracted from the carrying content file. However, in some cases, it can be perceptible. For example, it can be visibly added to a film or image in order to encourage the user to buy the original version without the mark.

The hidden information should be robust (resistant) against simple modifications normally used to the media type. For example, a marker hidden in an image should not be removed when the image is converted from Bitmap to JPG.

Steganographic techniques such as watermarking and steganography are evaluated using measures like robustness, imperceptibility, and capacity. In watermarking, the main objective is robustness. It should not be confused with steganography where imperceptibility is the main goal [45].

Figure 2 shows how the character "a" can be watermarked in a video using a simple watermarking scheme.
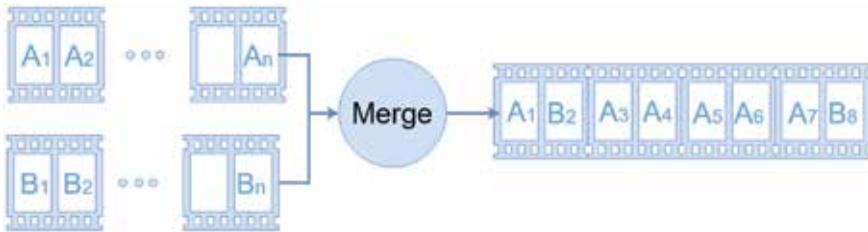
Figure 2.
Watermarking the character "a" in a movie.

In the example shown in **Figure 2**, the video is first decomposed into **n** segments. Then two different variants of the video represented by *A* and *B* are prepared which are imperceptibly different in each segment. Each segment can be chosen from the variant *A* or *B*. The choice of *A* for a segment watermarks the bit value "**0**," and choosing **B** means watermarking "1." In this example, the ASCII code "**01000001**" assigned to the character "*a*" has been watermarked to the video by choosing segments **2** and **8** from the variant **B** and the others from *A*.

4. Survey on watermarking in digital cinema

Illegal recording from the cinema screen using handheld cameras is a serious problem encountered by film makers. Watermarking, as a fingerprinting method aiming at piracy protection is a promising method for attacking this problem in the context of digital cinema. The watermarking process can be performed in show time and embed information regarding the projector and the playback time into the video. Since theatre owners are obliged to prevent handheld cameras in their premises, they are responsible for the pirate copies, and the watermark extracted from the pirate copies can specify the theatre where the copy has been recorded.

One of the most serious problems in the application of watermarking in digital cinema is that perspective, zoom, and some other parameters can be changed by the handheld camera, which leads to geometrical distortions. These distortions make it difficult to recover the watermark. This problem is faced by movie providers as they need to identify the theatre from which the pirate copy has been recorded. A watermarking system should guarantee distortion resistance in order to be capable of being used in digital cinema. In fact, old correlation-based watermark schemes are unable to prevent piracy because they are not robust to geometric distortions created by the handheld camera [46, 47].

The second challenge, which is faced by theatre owners, is locating the pirate in the theatre. Since theatre owners are sued when a pirate copy is distributed, they need mechanisms to identify the person who has recorded the illegal copy. To do this, they need to locate the camcorder. This way, they can identify the criminal using their ticket databases. This raises the need for watermarking methods capable of locating the camcorder.

The two aforementioned challenges have been part of the topic in several research works since early 2000s [48, 49], and they are still considered as research concerns in this area [42, 50]. In the following, we separately review the methods used by different watermarking schemes in order to attack each of the problems.

4.1 Identifying the theatre

In this section, we study the methods used in order to identify the theatres wherein hand camera copies are recorded. These methods mainly depend on geometric process while storing or detecting the watermark.

4.1.1 Temporal watermarking

One solution to the aforementioned problem is to design methods which use only the temporal access to embed the watermark. A watermarking scheme of this kind was proposed in [51, 52], which modifies the mean luminance value in each frame of the video according to the samples of the watermark. In this research report, it has been taken into consideration that the human visual system is sensitive to flickers in low spatial frequencies. Thus, the proposed scheme attempts at avoiding this kind of flickers. To do this, the same watermark is applied to a certain number of consequent frames.

In the aforementioned scheme, the watermark is considered as a pseudo-random sequence of length n consisting of 1 and -1 values. Embedding the watermark causes the luminance of each pixel to increase or decrease. But the change depends on a local scaling factor which has higher values for moving textual areas and lower values for non-moving flat areas. This scaling factor is defined as the minimum of a spatial scaling factor and a motion scaling factor. Moreover, there is a maximum allowable change that cannot be violated.

The motion scaling factor for each pixel is calculated as the difference from the corresponding pixel in the previous frame. On the other hand, the spatial scaling factor is calculated using a Laplacian filter. **Figure 3** shows this watermarking scheme.

In **Figure 3**, $T_{Texture}$ and $T_{Texture}$ are predefined thresholds and $|ABS|$ represents the absolute operator. Moreover, $0 \leq q < 1$, k is the number of the frame among the selected set of consecutive frames for embedding the watermark, and K is the number of consecutive frames.

The authors of this report successfully tested the robustness of their watermark under several conditions such as changing the zoom and angle of the camera in addition to handheld camera recording and people walking between the camera and the screen.

4.1.2 Compensation

Compensation is another solution to the geometrical distortion problem. In this method, a pre-process is used to restore the geometrical distortions before watermark extraction. A method based on compensation was proposed in [53]. The authors of this report made a simplifying assumption in order to make it easier to model the nonlinear distortions. They assumed that the handheld camera is located far from the screen at the back of the movie theatre but not so far from the center line of theatre.

The authors of [53] present a parameterized mathematical model that compares the pirate copy with its original in a general case. For each specific pair of pirate-original pair, the compensation method tries to find the values for the parameters such that the model best matches the case. They modeled the distortions with four mathematical transformations as follows.

- Affine transformations: there is assumed to be six degrees of parameters for this kind of transformations, which model rescaling (zooming), translation

(shifting caused by shakes of the camera), rotation (caused by the angle of the camera), and shearing (caused by cropping).

- Bilinear transform: this kind of transform can convert a square to an ordinary quadrilateral with eight degrees of freedom.

- Curved transform: tt is a simple mathematical transform that models the impacts of the lens properties such as short focal length.

- Curved-bilinear transform: it has been designed to compensate the bilinear transform as well as the curved transform.

They tested their method via simulating a handheld camera. The simulation process applies a curved-bilinear transformation along with converting from 1920 × 1980 resolution to 1024 × 576 and then cropping to 720 × 576. **Figure 4** shows this transformation. Their method successfully recovered a 64-bit watermark from the transformed video.
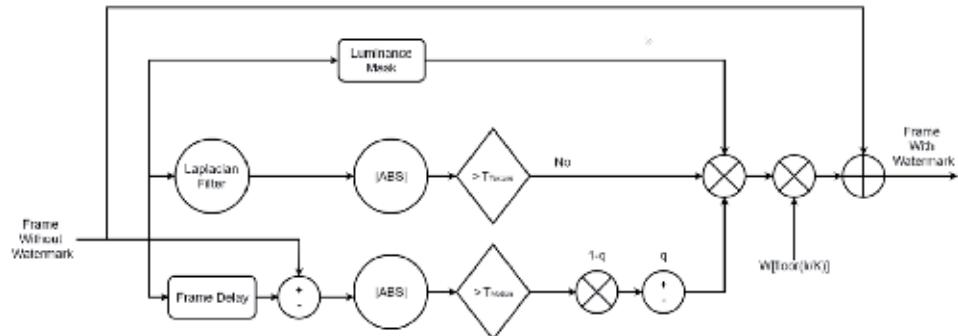
Figure 3.
The watermarking scheme presented in [51, 52].
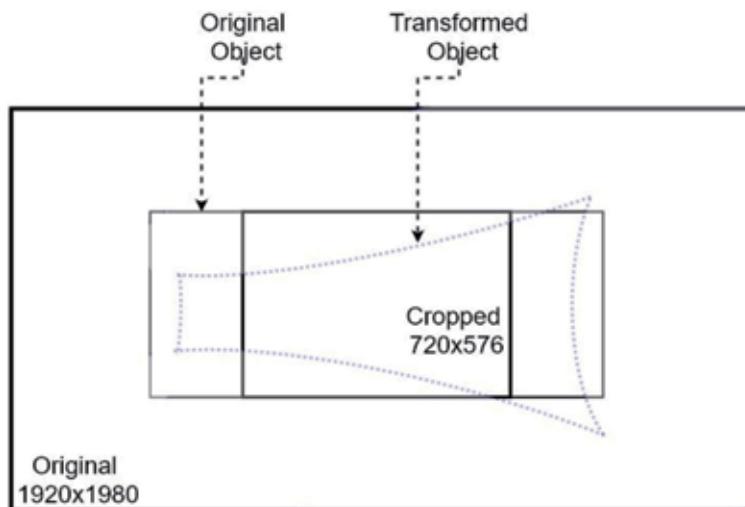
Figure 4.
The curved-bilinear transformation in [48].

4.1.3 Hardware-based additional watermarking

A hardware-based watermarking scheme named "Additional Watermarking" was presented in [54–56], which does not require the original movie for the recovery of the watermark. This method uses DWT (Discrete Wavelet Transform) instead of the Laplacian transforms. In this scheme, watermarks are added to the video in both the encoding and the decoding processes. This method uses CRC32 as a hash function for selecting random places in the video to embed the watermark. This method places the mark in the lowest frequency components of the video. It was shown in the research that the proposed method does not make any perceptible change in the original video and it is robust to a variety of signal processing transforms.

4.1.4 Spread transform dither modulation

The problem of tradeoff between fidelity and robustness was posed in [57]. A watermarking scheme based on STDM (Spread Transform Dither Modulation) was proposed in this research which uses the HVS (Human Visual System) properties of wavelet in order to resolve the tradeoff. These properties include the following facts.

- Noise in high resolution bands is not clearly perceptible to human eyes.

- HVS is less sensitive to changes to very high or very low brightness portions of the video.

- HVS is not very sensitive to changes in highly textured portions unless they are close to the borders.

In this scheme, the watermark is embedded during the JPEG2000 decoding phase. **Figure 5** shows a screen shot from Big Buck Bunny film before and after watermarking with this method.

4.1.5 Spread spectrum watermarking

In [53] another application for watermarking in digital cinema was proposed. This report proposes a watermarking method based on spread spectrum. This method is not only robust to geometric distortions, but also able to estimate the position of the handheld camera that has recorded the pirate copy. This may help identifying the ticket number and the credit card number of the person.
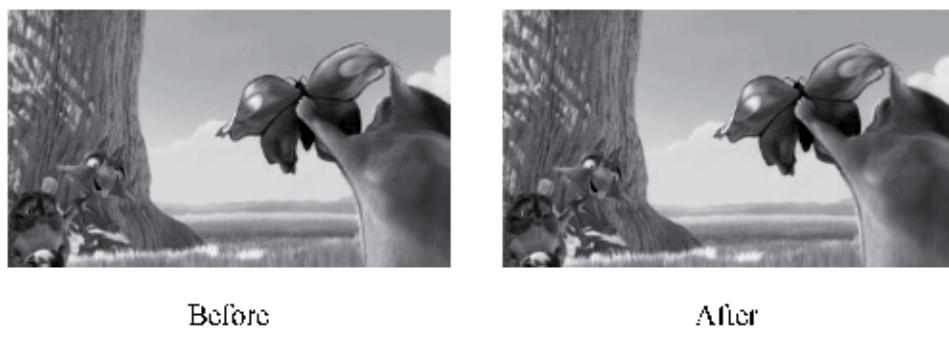
Before   After

Figure 5.
A screenshot from Big Buck Bunny before and after watermarking (Courtesy to [57]).

The method proposed in [58] embeds the watermark in the payload in a way that the geometric transformation applied by the handheld camera can be identified in addition to the position of the camera. This method requires the watermark payload to follow a periodic pattern. A basic pattern is generated using a secret key in a way that it follows a periodic Gaussian distribution with the mean equal to zero and the variance equal to one. Then the payload, which contains the serial number of the projector as well as a time stamp is embedded into the video after the modulation process using an additive spread-spectrum method.

The authors of [58] tested the accuracy of their method in a small seminar room as well as a large auditorium. The mean absolute error was at maximum 6.87 cm for the small-scale test and 50.38 cm for the large-scale test.

Another watermarking scheme based on spread spectrum techniques was introduced in [59]. This method stores the watermark in the YCbCr color space using translation, rotation, scaling or composite operations.

Table 1 compares the methods used to identify the theatre. The parameters used in this comparison are as follows.

- Resistance against hand vibrations

- Resistance against perspective drifts

- Resistance against obstacles such as people's heads or people walking between the handheld camera and the screen

- Resistance against zoom changes

One point to note here is that the distortion factors studied in **Table 1** are not the only factors that make it complex to recover the projection time watermark from pirate copies. For example, it was shown in [60] that the lamination flickers, induced by the interplay between an LCD and a camcorder can cause some difficulties in the recovery of the watermark. In addition to temporal luminance variations studied in [60], spatial luminance variations induced by camcorder recording have been studied in [61].

4.2 Locating the camcorder

In this section, we study the methods that can be used by cinema owners in order to locate the pirate camera.

		Method				
		Temporal	Compensation	Additional	Spread transform	Spread spectrum
Presented in		[51, 52]	[53]	[54–56]	[57]	[58, 59]
Robust Against	Vibration	NO	NO	YES	YES	NO
	Angle	YES	YES	YES	YES	YES
	Obstacles	YES	YES	YES	YES	YES
	Zoom	YES	YES	YES	YES	YES

Table 1.
A comparison among watermarking schemes used in digital cinema.

4.2.1 Audio watermarking

An approach to estimating the camcorder location has been proposed in [62]. In this method, unlike previous methods, audio is watermarked in the movie soundtrack instead of images, numbers or text messages being watermarked in the video itself. Most of the methods introduced in the literature avoid audio watermarking because of its complexity. In fact, it is difficult to hide any audio in soundtracks as they contain different types of audio such as voices, sound effects, and music.

The method proposed in [62] depends on a stochastic analysis in the watermark detection process. This method uses the audio watermarking technique introduced in [63]. The idea behind this method is that the attenuation of an audio signal depends on the distance. Thus, a few watermarked audio signals played from different locations can provide adequate information for locating the pirate. The advantage of this method is that audio is not prone to geometric distortion.

4.2.2 Hybrid audio-video watermarking

Some recent research works such as the one reported in [50] have combined video watermarking and audio watermarking in order to achieve the advantages of both methods. Audio watermarking is not affected by geometric distortions resulting from perspective, zoom, and vibrations. On the other hand, video watermarking is not affected by distance. Thus, it looks pertinent to combine them to achieve better results.

In the method proposed in [50], audio is watermarked in the video itself instead of the soundtrack. This way, there is no need for the watermarked audio to be played from different sources.

Another hybrid audio-video watermarking scheme was introduced in [64]. In this method unlike the one presented in [50], audio and video are separately watermarked in the soundtrack and in the movie itself.

5. Conclusion

In this chapter, we introduced the challenges faced by watermarking schemes in digital cinema. Then, we presented a review on watermarking methods specifically designed to confront these challenges. We focused on two main problems, first of which is the geometric distortion problem which and the second is camcorder location estimation. Our work in this chapter can be continued by studying other security and performance aspects of digital cinema.

Author details

Behrouz Zolfaghari[†] and Pinaki Mitra[*]
Department of Computer Science and Engineering, Indian Institute of Technology
Guwahati, India

*Address all correspondence to: pinaki@iitg.ac.in

† Equally contributed.

IntechOpen

References

[1] Hata M, Fukazawa A, Bessho M, Makino S, Higuchi M. A new 40 GHz digital distribution radio with single local oscillator. In: Proceedings of IEEE-MTT-S International Microwave Symposium Digest; 1978

[2] Hata M, Fukasawa A, Bessho M, Makino S, Higuchi M. A 40-GHz digital distribution radio with a single oscillator. IEEE Transactions on Microwave Theory and Techniques. 1980;**9**:951-962

[3] Stranathan BL. Issues to be considered—The implementation of a digital distribution system by the commercial broadcast networks. SMPTE Journal. 1997;**106**(11):795-798

[4] Rupp P, Estier T. A model for a better understanding of the digital distribution of music in a peer-to-peer environment. In: Proceedings of Annual Hawaii International Conference on System Sciences; 2003

[5] Zhao J, Liu H, Guangxian LV, Wang L, Ji Y. Research and development on digital distribution network. In: Proceedings of International Conference on Sustainable Power Generation and Supply; 2009

[6] Wang B, Huang T, Dong X, Jiang B, Bo ZQ. Dynamic emergency control in digital distribution substation with MW PV power. In: Proceedings of International Conference on Critical Infrastructure (CRIS); 2010

[7] Cilardo A, Barbareschi M, Mazzeo A. Secure distribution infrastructure for hardware digital contents. IET Computers and Digital Techniques. 2014;**8**(6):300-310

[8] Lee N-Y, Hwang T. On the security of Park et al.'s key distribution protocol for digital mobile communications. In: Proceedings of International Symposium on Personal, Indoor, and Mobile Communications (PIMRC); 1996

[9] Higuero M, Unzilla J, Saiz P, Jacob E, Aguado M, Goirizelaia I. A practical tool for analysis of security in systems for distribution of digital contents based on 'attack trees'. In: Proceedings of IEEE International Symposium on Broadband Multimedia Systems and Broadcasting; 2009

[10] Ghatak P, Tripathi RC, Chakravarti AK. Digital rights management: An integrated secure digital content distribution technology. Journal of Intellectual Property Rights. 2004;**9**:331-331

[11] Liu Q, Safavi-Naini R, Sheppard NP. Digital rights management for content distribution. In: Proceedings of the Australasian information security workshop conference on ACSW frontiers; 2003

[12] Gauche E, Coelho J, Treive RCG. An online distribution feeder optimal reconfiguration algorithm for resistive loss reduction using a multi-layer perceptron. In: Proceedings of International Conference on Neural Networks (ICNN); 1997

[13] Clemons EK, Aron R. Online distribution: A taxonomy of channel structures, determinants of outcome, and determinants of strategy. In: Proceedings of the 35th Annual Hawaii International Conference on System Sciences; 2002

[14] Azevedo DNR, De Oliveira JMP. Application of data mining techniques to the storage management and online distribution of satellite images. In: Proceedings of Seventh International Conference on Intelligent Systems Design and Applications (ISDA); 2007

[15] Lo SW. Towards secure online distribution of multimedia codestreams [thesis]. Singapore: Management University; 2016

[16] Kuang J, Yu S-Z. Broadcast-based content delivery in information-centric hybrid multihop wireless networks. IEEE Communications Letters. 2017;**21**(4):889-9892

[17] Xiao X, Ahmed M, Chen X, Zhao Y, Li Y, Han Z. Accelerating content delivery via efficient resource allocation for network coding aided D2D communications. IEEE Access. 2019;**7**:115783-115796

[18] Azogu IK, Ferreira MT, Liu H. A security metric for VANET content delivery. In: Proceedings of IEEE Global Communications Conference (GLOBECOM); 2012

[19] Chen M, Wang L, Chen J, Wei X, Lei L. A computing and content delivery network in the smart city: Scenario, framework, and analysis. IEEE Network. 2019;**33**(2):89-95

[20] Bilen T, Canberk B. Handover-aware content replication for mobile-CDN. IEEE Networking Letters. 2019;**1**(1):10-13

[21] Al-Abbasi A, Aggarwal V, Lan T, Xiang Y, Ra M-R, Chen Y-F. Fasttrack: Minimizing stalls for cdn-based over-the-top video streaming systems. IEEE Transactions on Cloud Computing. 2019;**27**(2):835-847

[22] Tang G, Wang H, Wu K, Guo D. Tapping the knowledge of dynamic traffic demands for optimal CDN design. IEEE/ACM Transactions on Networking. 2019;**27**(1):98-111

[23] Taleb T, Frangoudis PA, Benkacem I, Ksentini A. CDN slicing over a multi-domain edge cloud. IEEE Transactions on Mobile Computing (Early Access). 2019:1-1

[24] Cui S, Asghar MR, Russello G. Multi-CDN: Towards privacy in content delivery networks. IEEE Transactions on Dependable and Secure Computing (Early Access). 2018:1-1

[25] Chattopadhyay T, Sinha A, Hardikar A. H.264 compressed domain watermarking in content delivery network (CDN) environment. In: Proceedings of 2nd International Conference on Computational Intelligence, Communication Systems and Networks; 2010

[26] Fiadino P, D'Alconzo A, Bär A, Finamore A, Casas P. On the detection of network traffic anomalies in content delivery network services. In: Proceedings of 6th International Teletraffic Congress (ITC); 2014

[27] Li S, Doh I, Chae K. Non-redundant indirect trust search algorithm based on a cross-domain trust model in content delivery network. In: Proceedings of 19th International Conference on Advanced Communication Technology (ICACT); 2017

[28] Huang T-C, Shieh C-K, Miao Y-B. Java Application's Packet Eavesdropper for content delivery network. In: Proceedings of International Conference on Advanced Information Networking and Applications (AINA); 2005

[29] Boitard R, Jacquemin J-P, Damberg G, Stojmenovik G, Ballestad A. Evaluation of color pixel representations for high dynamic range digital cinema. SMPTE Motion Imaging Journal. 2018;**127**(2):46-56

[30] Gong B, Qin H, Chen D, Wang F. A study of stereoscopic digital cinema in china including new standards and recommendations. SMPTE Motion Imaging Journal. 2016;**125**(3):1-7

[31] Schuck M, Ludé P. An analysis of system contrast in digital cinema

auditoriums. SMPTE Motion Imaging Journal. 2016;**125**(4):40-49

[32] Stone JJ. A new integrated system for digital cinema projection and security. In: Proceedings of The IEE 2-Day Seminar on IT to HD; 2004

[33] Bloom JA. Digital Cinema Content Security and the DCI. In: Proceedings of Annual Conference on Information Sciences and Systems; 2006

[34] Stone JJ, David MWA. An integrated system for digital cinema projection and security. SMPTE Motion Imaging Journal. 2005;**114**:10-11

[35] Bloom JA. Security and rights management in digital cinema. In: Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP); 2003

[36] Bloom JA. Security and rights management in digital cinema. In: Proceedings of International Conference on Multimedia and Expo. (ICME); 2003

[37] Ma H, Jia C, Li S, Zheng W, Xmark W. Dynamic software watermarking using Collatz conjecture. IEEE Transactions on Information Forensics and Security. 2019;**14**(11):2859-2874

[38] Kricha Z, Kricha A, Sakly A. Accommodative extractor for QIM-based watermarking schemes. IET Image Processing. 2019;**13**(1):89-97

[39] Shayan M, Bhattacharjee S, Tang J, Chakrabarty K, Karri R. Bio-protocol watermarking on digital microfluidic biochips. IEEE Transactions on Information Forensics and Security. 2019;**14**(1):625-639

[40] Shukla D, Sharma M. A novel scene-based video watermarking scheme for copyright protection. Journal of Intelligent Systems. 2017;**27**(1):47-66

[41] Kelkoul H, Zaz Y. Digital cinema watermarking state of art and

comparison. International Journal of Computer, Electrical, Automation, Control and Information Engineering. 2017;**11**(2):256-261

[42] Dubey NK, Modi H. Comparatives study of various techniques against camcorder piracy in theater. In: 6th International Conference on Multimedia Computing and Systems (ICMCS); 2018

[43] Dubey NK, Kumar S. A review of watermarking application in digital cinema for piracy deterrence. In: ourth International Conference on Communication Systems and Network Technologies; 2014

[44] Bertrand N, Durou J-D, Charvillat V. The DCP bay: Toward an art-house content delivery network for digital cinema. In: Proceedings of 7th International Conference on Advances in Multimedia (MMEDIA); 2015

[45] Sharifzadeh M, Aloraini M, Schonfeld D. Adaptive batch size image merging steganography and quantized gaussian image steganography. IEEE Transactions on Information Forensics and Security. 2020;**15**:867-879

[46] Kalker T, Depovere D, Haitsma J, Maes M. A video watermarking system for broadcast monitoring. In: Proceedings of IS&T/SPIE/EI25; 1999

[47] Bender W, Gruhl D, Morimoto N. Techniques for data hiding. In: Proceedings of the SPIE; 1995

[48] Lubin J, Bloom JA, Cheng H. Robust content-dependent high-fidelity watermark for tracking in digital cinema. Proceedings of SPIE. 2003;**5020**(1):536-545

[49] Nguyen P, Balter R, Montfort N, Baudry S. Registration methods for nonblind watermark detection in digital cinema applications. Proceedings of SPIE. 2003;**5020**(1):553-562

[50] Kelkoul H, Zaz Y, Tribak H, Schaefer G. A robust combined audio and video watermark algorithm against cinema piracy. In: 6th International Conference on Multimedia Computing and Systems (ICMCS); 2018

[51] Kalker T, Haitsma J. A watermarking scheme for digital cinema. In: Proceedings of International Conference on Image Processing; 2001

[52] Van Leest A, Haitsma J, Kalker T. On digital cinema and watermarking. In: Proceedings of SPIE 5020; 2003

[53] Delannay D, Delaigle J-F, Barlaud M. Compensation of geometrical deformations for watermark extraction in the digital cinema application. In: Proceedings of SPIE Security and Watermarking of Multimedia Contents; 2003

[54] Vural S, Tomii H, Yamauchi H. Traceable robust watermarking for digital cinema system. In: Proceedings of International Symposium on Signal Processing and Its Applications; 2005

[55] Vural S, Tomii H, Yamauchi H. Video watermarking for digital cinema contents. In: Proceedings of European Signal Processing Conference; 2005

[56] Vural S, Tomii H, Yamauchi H. Robust digital cinema watermarking. International Journal of Computer and Information Engineering. 2008;**2**(10):3606-3611

[57] Darazi R, Callau P, Macq B. Secure and HVS-adaptive exhibition spread transform dither modulation watermarking for digital cinema. In: Proceedings of First IEEE International Workshop on Information Forensics and Security (WIFS); 2009:1-12

[58] Callau P, Darazi R, Macq B. Exhibition QIM-based watermarking for digital cinema. Proceedings of SPIE. 2009;**7254**(1):1-12

[59] Dubey NK, Kumar S. An effective approach of distortion-resistant video watermarking for piracy deterrence. International Journal of Security and Its Applications. 2015;**9**(1):283-294

[60] Hajj-Ahmad A, Baudry S, Chupeau B, Doerr G, Wu M. Flicker forensics for camcorder piracy. IEEE Transactions on Information Forensics and Security. 2017;**12**(1)

[61] Kelkoul H, Zaz Y. Registration methods for nonblind watermark detection in digital cinema applications. International Journal of Computer, Electrical, Automation, Control and Information Engineering. 2003;**11**(2):2107

[62] Nakashima Y, Tachibana R, Babaguchi N. Watermarked movie soundtrack finds the position of the camcorder in a theater. IEEE Transactions on Multimedia. 2009;**11**(3)

[63] Tachibana R, Shimizu S, Kobayashi S, Nakamura T. An audio watermarking method using a two-dimensional psuedo-random array. Signal Processing. 2002;**82**(1):1455-1469

[64] Gosavi CS, Mali SN. Secure, robust video watermarking to prevent camcorder piracy. Indian Journal of Science and Technology. 2017;**10**(18):1-10

Chapter 3

Confidentiality and Integrity for IoT/Mobile Networks

Tri Ngo Minh

Abstract

This chapter discusses how to ensure confidentiality and integrity for data flow in IoT applications. While confidentiality could be assessed by access control, cryptography, or information flow analysis, integrity is still an open challenge. This chapter proposes to use error-correcting codes to guarantee integrity, i.e., to maintain and assure the errorless state of data. Besides errors, many communication channels also cause erasures, i.e., the receiver cannot decide which symbol the received waveform represents. The chapter proposes a method that might correct both errors and erasures together. Our method is efficient in reducing memory storage as well as decoding complexity.

Keywords: confidentiality, integrity, information flow, erasure, separating matrix, covering design

1. Introduction

It is estimated that Internet of Things (IoT) will generate billions of dollars in profit for industries over the next two decades. Many organizations have started to develop and implement their own IoT strategies. IoT enables devices would generate and transmit so many data such that security should be a top concern. IoT users require that communication technologies have to guarantee both efficiency and security. This chapter discusses how to guarantee two main properties of security, i.e., *confidentiality* and *integrity*, for IoT applications.

1.1 Confidentiality

Securing the data manipulated by information systems has been a challenge in the past few years. Several methods to limit the information disclosure have been proposed, such as *access control* and *cryptography*. These are useful approaches, i.e., they can prevent confidential information from being read or modified by *unauthorized* users. However, they still have a fundamental limitation, i.e., they do not regulate the information propagation after it has been released. For example, access control prevents unauthorized file access, but is insufficient to control how the data is used afterwards. Similarly, cryptography provides a shield to exchange information privately across a nonsecure channel, but no guarantee about the confidentiality of private data is given after it is decrypted. Thus, neither access control nor encryption provides a *complete* solution to protect confidentiality for information systems.

To ensure confidentiality for an information system, i.e., IoT system, it is necessary to show that the system *as a whole* enforces a confidentiality policy, i.e., by analysing how information flows within the system. The analysis must show that information controlled by a confidentiality policy cannot flow to a place where that policy is violated. Thus, the confidentiality policy we wish to enforce is an *information flow policy*, and the method that enforces them is an *information flow analysis*.

Information flow analysis is a technique that has recently become an active research topic. In general, the approach of information flow security is based on the notion of *interference* [1]. Informally, *interference* exists inside a system when private data affect public data, e.g., an attacker might guess private data from observing public data. *Noninterference*, i.e., the absence of interference, is often used to prove that an information system is secured.

Noninterference is required for applications where the users need their private data strictly protected. However, many practical IoT applications might leak *minor* information. Such systems include password checkers, cryptographic operations, etc. For instance, when an attacker tries to guess the password: even when the attacker makes a wrong guess, secret information has been leaked, i.e., it reveals information about what the real password is not. Similarly, there is a flow of information from the plain-text to the cipher-text, since the cipher-text depends on the plain-text. These applications are rejected by the definition of noninterference.

However, the insecure property will happen only when it exceeds a specific threshold, or amount of interference. If the interference in the system is small enough, e.g., below a threshold given by specific security policy, the system is considered to be secure. The security analysis that requires to determine how much information flows from high level, i.e., secret data, to low level, i.e., public output, is known as *quantitative information flow*. It concerned with measure the leakage of information in order to decide if the leakage is tolerable.

Qualitative information flow analysis, i.e., noninterference, aims to determine whether a program leaks private information or not. Thus, these absolute security properties always reject a program if it leaks any information. *Quantitative* information flow analysis offers a more general security policy, since it gives a method to tolerate a minor leakage, i.e., by computing how much information has been leaked and comparing this with a threshold. By adjusting the threshold, the security policy can be applied for different applications, and in particular, if the threshold is 0, the quantitative policy is seen as a qualitative one. The idea of quantitative information flow analysis has been discussed in details in [2], one of our papers; readers can refer to it for more information.

1.2 Integrity

Integrity means maintaining and assuring accuracy and completeness of data. However, during the wireless transmission in IoT applications, messages can be erroneous due to many reasons, e.g., attenuation, distortion or the addition of noise. Error means the receiver cannot decode correctly the signal to get the right symbol. In order to protect data against errors, channel coding, i.e., error-correcting codes are required. Error-correcting codes ensure proper performance of IoT systems. They ensure the integrity of communication links in the presence of noise, distortion, and attenuation [3–6]. The use of a parity-bit as an error-detecting mechanism is one of the simplest and most well-known schemes used in digital communication. Data is portioned into blocks. To each block, an additional bit is appended to make the number of bits which are 1 in the block, including the appended bit, an even number. If a single bit-error occurs, within the block, the number of 1's becomes odd. Hence, this allows for detection of single errors [7, 8].

Error-correcting codes are often applied in telecommunications. Many early applications of coding were developed for deep-space and satellite communication systems. For example, signals from satellites and space crafts are sent back to earth. The channel for such transmission is space and the earth's atmosphere. These communication systems not only have limitations on their transmitted power, but also introduce errors, due to solar activity and atmospheric conditions, into weak signals. Error-correcting codes are an excellent method to guarantee the integrity of these communication links. With the applications of error-correcting codes, most of the data sent could be correctly decoded here on earth. As examples, a binary (32,6,16) Reed-Muller code was used during the Mariner and Viking mission to Mars around 1970 or a convolutional code was used on the Pioneer 10 and 11 missions to Jupiter and Saturn in 1972. The (24,12,8) Golay code was used in the Voyager 1 and Voyager 2 spacecrafts transmitting color pictures of Jupiter and Saturn in 1979 and 1980. When Voyager 2 went on to Uranus and Neptune, the code was switched to a concatenated Reed-Solomon code for its substantially more powerful error correcting capabilities.

The block and convolutional codes are also applied to the global system for mobile communications (GSM) which is the most popular digital cellular mobile communication system. Reed Solomon and Viterbi codes have been used for nearly 20 years for the delivery of digital satellite TV. Low-density parity-check codes (LDPC codes) are now used in many recent high-speed communication standards, such as Digital video broadcasting-S2 (DVB-S2), WiMAX, 10GBase-T Ethernet [9].

Most error correcting codes, in general, are designed to correct or detect errors. However, many channels cause erasures, i.e., the demodulator cannot decide whether the received waveform represents bit 0 or 1, in addition to errors. Basically, decoding over such channels can be done by: firstly, deleting erased symbols and then, decoding the resulting vector with respect to the punctured code, i.e., the code in which all erasures have been removed. For any given linear code and any given maximum number of correctable erasures, in [7], Abdel-Ghaffar and Weber introduced a parity-check matrix yielding parity-check equations that do not check any of the erased symbols and which are sufficient to characterize the punctured code. This allows for the *separation* of erasures from errors to facilitate decoding. However, these parity-check matrices have too many redundant rows. To reduce decoding complexity, parity-check matrices with small number of rows are preferred. This chapter proposes a method that can build a matrix with a smaller number of rows.

Organization of the paper: The rest of this chapter is organized as follows. Section 2 introduces the main ideas of error-correcting codes, errors and erasures. Section 3 presents methods to construct a parity-check matrix that can correct both errors and erasures. Section 4 discusses a general solution for the covering design, which is used in the proposal. Finally, Section 5 concludes the chapter.

2. Codes, errors and erasures

2.1 Linear block codes

Let C be an $[n, k, d]$ linear block code. It means that C is a k-dimensional subspace of the n-dimensional vector space. The set of codewords of C can be defined as the null space of the row space of an $r \times n$ parity-check matrix $H = (h_{i,j})$ of rank $n - k$. Since a vector $\mathbf{x}$ is a codeword of C iff $\mathbf{x}H^T = 0$, where the superscript T denotes the transpose, we can derive r parity-check equations PCE, as follows,

$$\text{PCE}i : \sum_{j=1}^{n} h_{i,j} x_j = 0 \text{ for } i = 1, 2, ..., r. \tag{1}$$

An equation $\text{PCE}i(\mathbf{x})$ is said to check $\mathbf{x}$ in position j iff $h_{i,j} \neq 0$.

2.2 Erasures

Sometimes, at the receiver, the demodulator cannot decide which symbol the received waveform represents. In this case, we declare the received symbol as an *erasure*. When the received codeword contains erasures instead of errors, the iterative decoding can be used [8].

Here, we summarize the iterative decoding procedure using an example of the (7,4,3) binary Hamming code with the following parity-check matrix,

$$H = \begin{pmatrix} 1 & 0 & 1 & 1 & 1 & 0 & 0 \\ 0 & 1 & 0 & 1 & 1 & 1 & 0 \\ 0 & 0 & 1 & 0 & 1 & 1 & 1 \end{pmatrix}$$

Since a vector $\mathbf{x} = (x_1 x_2 x_3 x_4 x_5 x_6 x_7)$ is a codeword iff $\mathbf{x}H^T = 0$. Hence, every codeword has to satisfy three parity-check equations as follows.

Assume that the received vector is $*010*0$, where the erased symbol is denoted by Equation A checks on x_1, x_3, x_4 and x_5. If exactly one of these four symbols is erased, it can be retrieved from this equation. Thus, $x_1 = 1$ since $x_3 = 0$, $x_4 = 1$, and $x_5 = 0$. Similarly, we can derive that $x_2 = 1$, and $x_6 = 0$ from Equation B and C. Therefore, the iterative decoding decided that the transmitted codeword is 1101000.

Iterative decoding is successful iff erasures do not fill the positions of a *nonempty stopping set*. A stopping set is a set of positions in which there is no parity-check equation that checks exactly one symbol in these positions. The performance of iterative decoding techniques for correcting erasures depends on the sizes of the stopping sets associated with the parity-check matrix representing the code. The parity-check matrix with redundant rows could benefit the decoding performance, i.e., reducing the size of stopping sets, while increasing the decoding complexity. More information on stopping set can be found in [2, 10, 11].

2.3 Separation of errors from erasures

In this part, we discuss how to handle errors together with erasures. In this case, we can apply an algorithm using trials in which erasures are replaced by 0 or 1; and the resulting vector is decoded by a decoder which is capable of correcting errors. For binary code, two trials are sufficient [8, 12].

For example, if C is a binary (n, k)-code with a Hamming distance $d = 2t_{\varepsilon} + t_? + 1$, then C can correct t_{ε} errors and $t_?$ erasures. In the presence of *no* erasures, C is able to correct up to $t_{\varepsilon} + \lfloor t_?/2 \rfloor$ errors. Let $\mathbf{r}$ be a received vector having at most t_{ε} errors and at most $t_?$ erasures. Suppose the decoder constructs two vectors $\mathbf{r}0$ and $\mathbf{r}1$, where $\mathbf{r}i$ is obtained by filling all erasure positions in $\mathbf{r}$ with the symbols i, $i = 0, 1$. Since C is binary, in either $\mathbf{r}0$ or $\mathbf{r}1$, at least half of the erasure locations has the right symbols. Hence, either $\mathbf{r}0$ or $\mathbf{r}1$ has a distance at most $t_{\varepsilon} + \lfloor t_?/2 \rfloor$ from the transmitted codeword. Thus, any standard error correction technique can be applied. If the correction decodes both $\mathbf{r}0$ and $\mathbf{r}1$ to codewords,

and these codewords are the same, then this is the transmitted codeword. If they are different, then there is one, and only one, vector requiring at most t_ε changes in nonerasure positions to become the right codeword. More information on this algorithm can be found in [6].

Abdel-Ghaffar and Weber proposed another way of decoding over such channels [7]. First, all erasures are deleted from the received message. Errors in the resulting codeword will be corrected based on the punctured code, i.e., codewords consist of symbols in positions which are not erased. After all errors have been corrected, erasures will be recovered by the iterative decoding.

The decoder can compute a parity-check matrix for the punctured code after receiving the codeword. However, this leads to time delay which is unacceptable specially in IoT applications. To reduce time delay, we can store parity-check matrices of all punctured codes corresponding to all erasure patterns. The drawback of this solution is the requirement of huge memory storage at the decoder.

Abdel-Ghaffar and Weber proposed using a *separating matrix* with redundant rows, providing enough parity-check equations which do not check any of the erased symbols and are sufficient to form a parity-check matrix for the punctured code obtained by deleting all erasures [7]. Having these parity-check equations not checking any of the erased symbols lead to the concept of *separation* of errors from erasures.

The basic concept of this decoding technique can be illustrated by an example as follows. We consider an (8,4,4) binary extended Hamming code with the following parity-check matrix, **Figure 1**.

A normal parity-check matrix just has only four rows as the first four rows in this separating matrix. Allowing redundant rows simplifies the decoding of erasures in addition to errors. Assume that we get a codeword **r** = 0 ∗ 011000 with one erasure in the second position. Applying the decoding technique mentioned above, firstly we delete the erasure and the resulting vector is **r'** = 0011000. This vector **r'** can be considered as a codeword of the (7,4,3) punctured code. In H, the first, the second and the sixth row have zeros in the second position. It means that three corresponding parity-check equations do not check the erased symbol. Based on these three rows, we can form a parity-check matrix H' for the punctured code, as follows **Figure 2**.

Using H', **r'** is decoded into 0011010. Putting back the erasure, we get 0*011010. The third row of H, which checks the erased symbol, can be used to recover the erasure. Thus, the decoded codeword corresponding to **r** is 01011010.

A normal parity-check matrix cannot be used for decoding of both errors and erasures together. Decoding is feasible when we pay the price of storing a parity-check matrix with more rows than a normal one. In order to reduce the memory

$$H = \begin{pmatrix} 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 \\ 0 & 0 & 1 & 1 & 0 & 0 & 1 & 1 \\ 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 \\ 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 \end{pmatrix}$$

Figure 1.
A parity check matrix for the code C.

Figure 2.
A parity check matrix for the punctured code.

storage as well as the decoding complexity, a parity-check matrix with small number of rows is preferred.

Given any linear code and any given maximum number of correctable erasures, Abdel-Ghaffar and Weber introduced separating matrices yielding parity-check equations that do not check any of the erased symbols and which are sufficient to characterize all punctured codes corresponding to this maximum number of erasures [7]. This allows for the separation of erasures from errors to facilitate decoding. However, their proposal yields separating matrices which typically have *too many* redundant rows. The following part of this chapter discusses an improved method to construct such separating matrices, applying covering design, with a *smaller number of rows*.

3. How to build an *l*-separating matrix

3.1 Set separation

Let $H = (h_{i,j})$ of rank $n - k$ be an $(r \times n)$ parity-check matrix of C, $r \geq n - k$. Let S be a subset of $\{1, 2, ..., n\}$ and T be a subset of $\{1, 2, ..., r\}$, define $H_S^T = (h_{i,j})$ with $i \in T$ and $j \in S$, be a $T \vee \times \vee S \vee$ submatrix of H. For the code C with the length n, define $C_{\acute{S}} = \{c_{\acute{S}} : c \in C\}$ be the punctured code consisting of all codewords of C in which the symbols in positions indexed by S, $S = \{1, 2, ..., n\}\{S$ are deleted. Clearly, $C_{\acute{S}}$ is a linear code over $GF(q)$ of length $n' = \acute{S}\vee$, dimension $k' \leq k$, and Hamming distance $d' \leq d - \acute{S}\vee$. Let $\widetilde{S} = \{i : 1 \leq i \leq r, h_{ij} = 0 \forall j \in S\}$, define $H(S) = H_{\acute{S}}^{\widetilde{S}}$.

Definition 1 [7]: A parity-check matrix H separates $S \subseteq \{1, 2, ..., n\}$ iff $H(S)$ is a parity-check matrix of $C_{\acute{S}}$.

Theorem 1 [7]: A parity-check matrix H of an $[n, k, d]$ linear code C separates a set S of size $|S| \leq d - 1$ iff $H(S)$ has rank $n - k - |S|$.

Definition 2 [7]: If H separates all sets S of size l for a fixed $l \leq \min\{d, n - k\} - 1$, it is *l*-separating.

If H is an *l*-separating parity-check matrix of the code C, based on H, we can construct a parity-check matrix for any code punctured up to a fixed number l of symbols. H has two features:

- H can separate erasures from errors, since H has enough parity-check equations that do not check any erased symbols, and are sufficient to characterize the punctured code. It means that the punctured code, which is formed by deleting erased symbols, can be corrected errors by a sub-matrix of H.

- In case $l \leq \min\{d, n-k\} - 1$, H has no stopping set of size l or less. For any pattern of l or fewer erasures, not only are there enough parity-check equations that do not check any of the erased symbols characterize the punctured code, but also there is a parity-check equation that checks exactly one of the erased symbols. It means that after all errors have been corrected, erasures can be recovered by the iterative decoding procedure.

3.2 Separating matrix

Let H' be a full rank parity-check matrix, $S_i \subseteq \{1, 2, ..., n\}$, in which $i = 1, 2, ..., \binom{n}{l}$, be distinct subsets of $\{1, 2, ..., n\}$ of size l, For each i, it is trivial that H'_{S_i} has rank l ($l \leq \min\{d, n-k\} - 1$). By elementary row operations on H', we can obtain an $(n-k) \times n$ matrix H'_i, for each $i = 1, 2, ..., \binom{n}{l}$, of rank $n-k$, such that its last $n-k-l$ rows have zeros in positions indexed by S_i **Figure 3**.

Let H_I be a matrix which rows is the union of sets of the last $n-k-l$ rows in H'_i, for $i = 1, 2, ..., \binom{n}{l}$. H_I is an l-separating matrix of the code C, and it has at most $\binom{n}{l}(n-k-l)$ rows [7] **Figure 4**.
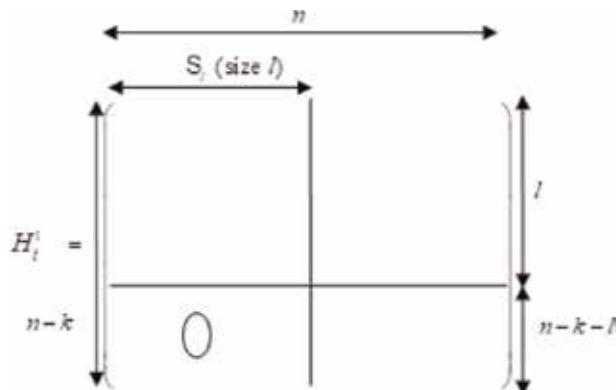
Figure 3.
Independent-row separation.

Figure 4.
An l-separating matrix.

3.3 A more efficient separating matrix

In this section, we propose a method that can construct an l-separating matrix with a smaller number of rows. This method implements the idea of *covering design* [13, 14]. Basically, given $1 \leq t \leq u \leq v$, a (v, u, t) covering design is a collection of u-element subsets of $V = \{1, 2, ..., v\}$, called blocks, such that each t-element subset of V is contained in at least one block, e.g., $\{1, 2\}$ is contained in $\{1, 2, 3\}$.

For our specific situation, consider an (n, b, l) covering design. Let $B = \{B_i\}$ be a set of b-element subsets, $1 \leq l \leq b \leq \min\{d, n - k\} - 1$, such that every l-element subset S_i is contained in at least one member of B. Assign to each S_i,

$i = 1, 2, ..., \binom{n}{l}$, an element B_j of B such that S_i is contained in B_j. H'_{B_j} has rank b.

For any B_j, by elementary row operations on H', we can obtain an $(n - k) \times n$-matrix of rank $n - k$ such that its last $n - k - b$ rows have zeros in positions indexed by B_j. After arranging columns, we obtain a matrix H_j^1 with the following format (Step 1) **Figure 5**.
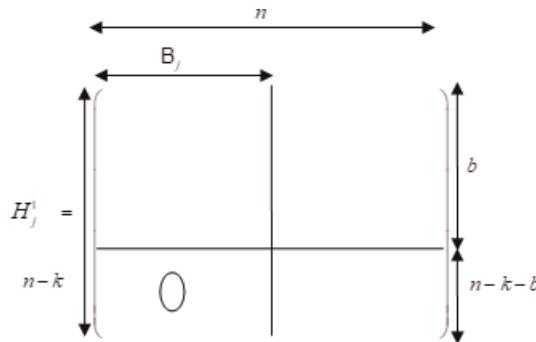
Figure 5.
Row separation—Step 1.
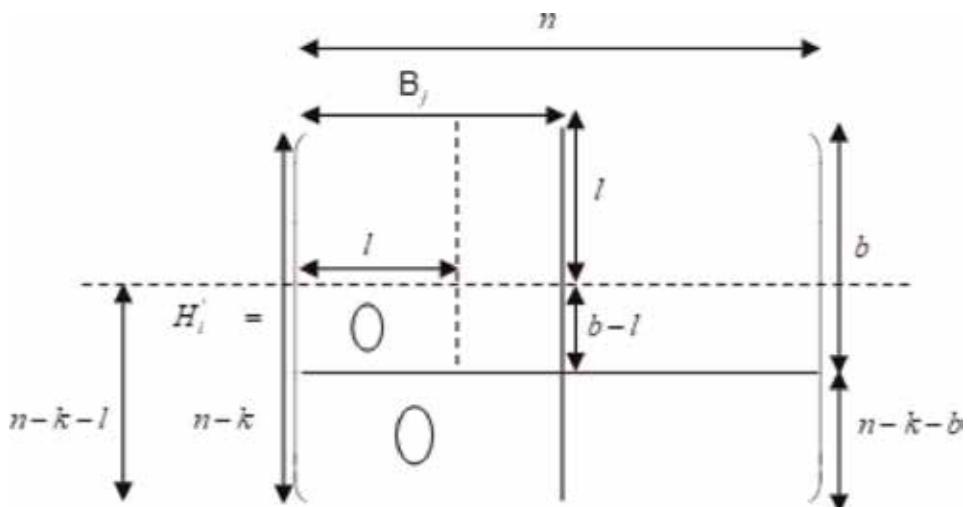
Figure 6.
Row separation—Step 2.
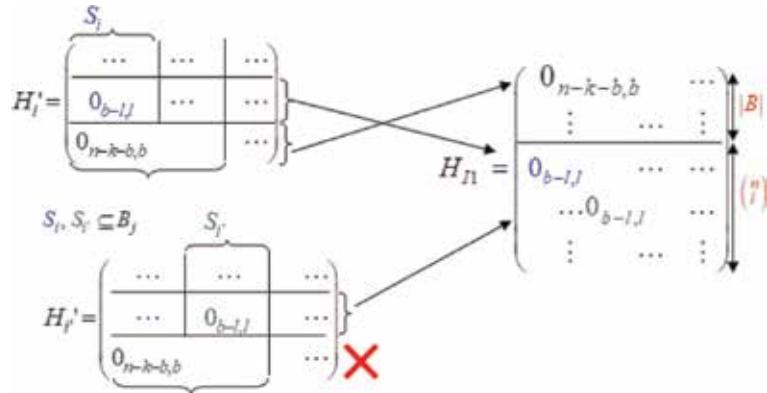
Figure 7.
A more efficient l-*separating matrix.*

Consider the set S_i assigned to B_j, by further elementary row operations, H_j^1 can be changed into a matrix such that rows $l + 1, l + 2, ..., b$ have zeros in positions indexed by S_i, and rows $b + 1, b + 2, ..., n - k$ have zeros in positions indexed by B_j. After column arrangement, we obtain a matrix with the following format (Step 2), **Figure 6**.

Following this method, if S_i and $S_{i'}$ belong to the same B_j, the last $n - k - b$ rows in H'_i and $H'_{i'}$ are the same. It follows that the matrix which rows is the union of the last $n - k - l$ rows in $H'_j, j = 1, 2, ..., \vee B \vee$, and the rows $l + 1, l + 2, ..., b$ of H'_i, $i = 1, 2, ..., \binom{n}{l}$, is an l-separating parity-check matrix of C. Let $B(n, b, l)$ denote the *minimum* size of B, i.e., $B(n, b, l) = \min |B|$. This matrix has at most $(n - k - b)B(n, b, l) + \binom{n}{l}(b - l)$ rows. It is obvious to see that the upper bound on number of rows in Approach 2 is strictly smaller than in Approach 1. In case $b = l$, two approaches are the same. For a given l, we can choose an appropriate b to achieve the best result **Figure 7**.

4. Covering design

Consider a (v, u, t) covering design, where $1 \leq t \leq u \leq v$.

Example 1: Given that $v = 8, u = 3, t = 2$. There are $\binom{8}{2}$ = 28 subsets of 2-elements, and $\binom{8}{3}$ = 56 subsets of 3-elements of $V = \{1, 2, ..., 8\}$. However, we only need at most 21 subsets of 3-elements, i.e., {{1,2,3}, {1,2,4}, {1,2,5}, {1,2,6}, {1,2,7}, {1,2,8}, {1,3,4}, {1,3,5}, {1,3,6}, {1,3,7}, {1,3,8}, {1,4,5}, {1,4,6}, {1,4,7}, {1,4,8}, {1,5,6}, {1,5,7}, {1,5,8}, {1,6,7}, {1,6,8}, {1,7,8}}, to form all 28 subsets of 2-elements. For example, based on the subset {1,2,3}, we can form {1,2}, {2,3}, {1,3}. Using 21 subsets of size 3 mentioned above, we can construct all 28 subsets of size 2.

The covering design problem has been investigated since many years ago. However, until now, there is no general *optimal* solution for all triples (v, u, t). In this section, we propose a *covering design* valid for all triples (v, u, t). This design is not optimal but it can give a general solution for the problem.

4.1 Approach 1

Firstly, we show that with at most $\binom{v-(u-t)}{t}$ subsets of size u, we can form all $\binom{v}{t}$ subsets of size t.

1. Take the first $u - t$ elements, i.e., $\{1, 2, ..u - t\}$, out of $V = \{1, 2, ..., v\}$.

2. The rest of the set is $\{u - t + 1, u - t + 2, ..., v - 1, v\}$. Based on these elements, form all subsets of size t. The number of subsets is $\binom{v-(u-t)}{t}$.

3. Put the first $u - t$ elements into each subset of size t to have subsets of size u. With these $\binom{v-(u-t)}{t}$ subsets of size u, it is easy to see that we can form all $\binom{v}{t}$ subsets of size t.

4.2 Approach 2

By modifying *Approach 1*, we show that some u-element subsets can be *merged* to reduce B_V.

1. Take the first $u - t$ elements, i.e., $\{1, 2, ..u - t\}$, out of $V = \{1, 2, ..., v\}$.

2. The rest of the set is $\{u - t + 1, u - t + 2, ..., v - 1, v\}$. Based on these elements, form all subsets of size t and arrange them into columns based on the following rules:

   - Elements in each subset are arranged in ascending order, e.g., {1,2,3}.

   - Subsets are arranged into columns. Subsets are in one column iff their first $t - 1$ elements are the same (except the *special column* mentioned below). Hence, subsets in one column are different from each other only in the last element. The subset with the smaller last element will be listed above.

   - *Special column*: In case $t \geq 2$, we arrange all subsets containing both element $v - 1, v$ in a column and name it special column. It is easy to see that there are $\binom{v-(u-t)-2}{t-2}$ subsets in this column.

3. Put the first $u - t$ elements into each subset of size t to have subsets of size u.

4. If the number of subsets in the *longest column* is greater or equal to three and the *special column* exists, we can merge the last two subsets, which contain either $v - 1$ or v, in each column (except the *special column*) into one, i.e., the *merged set*, by this rule

   - Take the union of two subsets, i.e., the size of the union subset is $u + 1$.

   - Eliminate the first element of the union subset, i.e., its size is now u.

the *longest column*   the *special column*

12345	12356	12367	12378	12389
12346	12357	12368	12379	12489
12347	12358	12369		12589
12348	12359		12478	12689
12349		12467	12479	12789
		12468		13489
	12456	12469	12578	13589
	12457		12579	13689
	12458	12567		13789
	12459	12568	12678	14589
		12569	12679	14689
				14789
		13467	13478	15689
		13468	13479	15789
		13469		16789
			13578	
		13567	13579	
		13568		
	13456	13569	13678	
	13457		13679	
	13458			
	13459		14578	
			14579	
		14567		
		14568	14678	
		14569	14679	
			15678	
			15679	

Figure 8.
The first three steps of Approach 2.

We can merge the last two subsets in each column (except the special column) because: (a) the first $u - 1$ elements in the last two subsets are also in another subset in the columns. Thus, any subset of size t formed by using these $u - 1$ elements can be form by any other subset in the column; (b) any subset of size t containing $\{1, v - 1\}$ or $\{1, v\}$ can be formed by subsets in the *special column*.

Example 2: Given that $v = 9, u = 5, t = 4$. Following the first three steps of Approach 2, we get:

First, take $\{1\}$ out of the set. Following Step 2, form all subsets of size 4, and arrange them into columns. Put $\{1\}$ back into each subset of size 4 to have subsets of size 5. We denote subsets in boxes are subsets which can be merged **Figure 8**.

The merged step (Step 4): For example, consider the first column, two subsets in the box are $\{1, 2, 3, 4, 8\}$ and $\{1, 2, 3, 4, 9\}$. First, take the union of the two, i.e., $\{1, 2, 3, 4, 8, 9\}$, and then eliminate the first element; thus, this results in $\{2, 3, 4, 8, 9\}$.

Therefore, Step 4 of Approach 2 gives the following result. It is easy to see that any subset of size 4 can be formed by subsets in **Figure 9**.
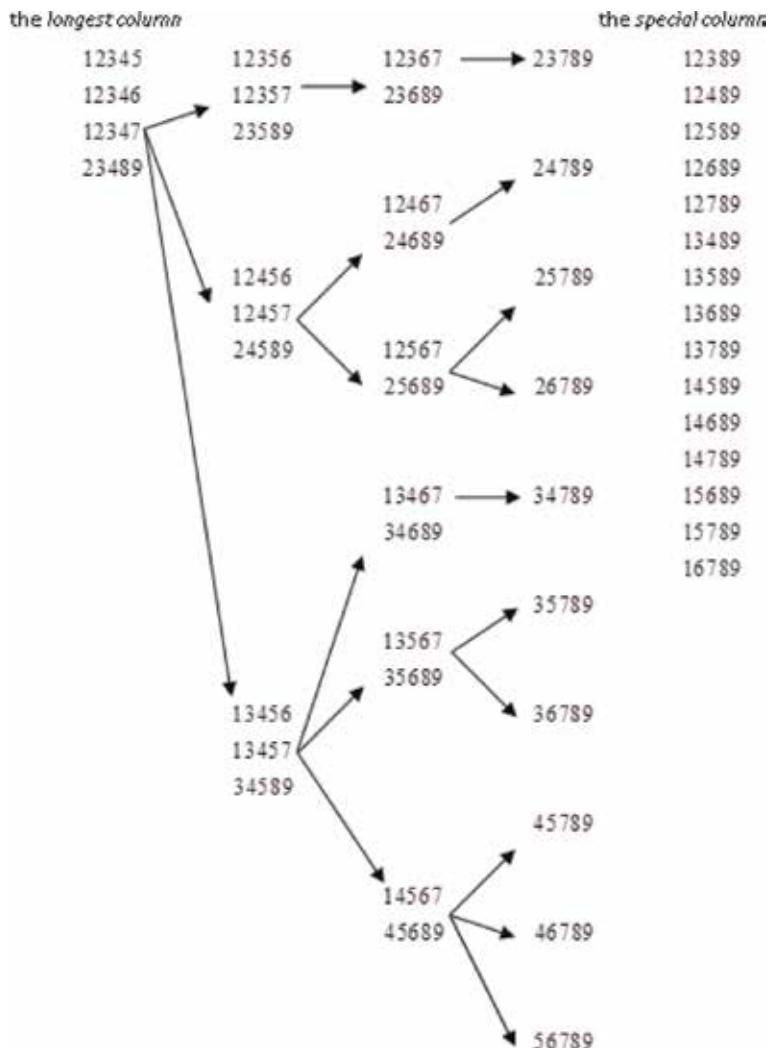
Figure 9.
The last step of Approach 2.

The number of reduced subsets is equal to the number of subsets that contain the elements $v - 1$ or v. Thus, the number of reduced subsets is $\binom{v - (u - t) - 2}{t - 1}$. Therefore, with at most $\binom{v - (u - t)}{t} - \binom{v - (u - t) - 2}{t - 1}$ subsets of size u, we can form all $\binom{v}{t}$ subsets of size t.

5. Conclusions

This chapter discusses how to ensure confidentiality and integrity for data in IoT systems. The chapter focuses more on integrity which can be ensured via the implementation of error-correcting codes. Separating parity-check matrices are useful for decoding over channels causing both errors and erasures. We propose a

way to build a separating parity-check matrix with a smaller set of rows. This method reduces both decoding complexity and memory storage. Besides, we also present a covering design. This design is not optimal but it gives a general solution for all triple (v, u, t).

Acknowledgements

Author details

Tri Ngo Minh
Faculty of Electronic and Telecommunication Engineering, The University of Danang—University of Science and Technology, Vietnam

*Address all correspondence to: tringominh@gmail.com

IntechOpen

References

[1] Goguen JA, Meseguer J. Security policies and security models. In: IEEE Symposium on Security and Privacy; 1982. pp. 11-20

[2] Ngo TM, Huisman M. Complexity and information flow analysis for multi-threaded programs. The European Physical Journal Special Topics. 2017;**226**(10):2375-2392

[3] Roth RM. Introduction to Coding Theory. Cambridge, UK: Cambridge University Press; 2006

[4] Lin S, Costello DJ Jr. Error Control Coding. Pearson Education International; 2004

[5] Mac Williams FJ, Sloane NJA. The Theory of Error-Correcting Codes. North-Holland Publishing Company; 1977

[6] Vanstone SA, van Oorschot PC. An Introduction to Error-Correcting Codes with Applications. Norwell, MA: Kluwer; 1989

[7] Abdel-Ghaffar KAS, Weber JH. Separating erasures from errors for decoding. In: Proceedings of the IEEE International Symposium on Information Theory; Toronto, Canada; 2008. pp. 215-219

[8] Weber JH. Lecture Notes: Error-Correcting Codes. Delft University of Technology; 2007

[9] Di C, Proietti D, Telatar IE, Richardson TJ, Urbanke RL. Finite-length analysis of low-density parity-check codes on the binary erasure channel. IEEE Transactions on Information Theory. 2002;**48**(6): 1570-1579

[10] Schwartz M, Vardy A. On the stopping distance and the stopping redundancy of codes. IEEE Transactions on Information Theory. 2006;**52**(3): 922-932

[11] Weber JH, Abdel-Ghaffar KAS. Results on parity-check matrices with optimal stopping and/or dead-end set enumerators. IEEE Transactions on Information Theory. 2008;**54**(3): 1368-1374

[12] Hollmann HDL, Tolhuizen LMGM. On parity check collections for iterative erasure decoding that correct all correctable erasure patterns of a given size. IEEE Transactions on Information Theory. 2007;**53**(2):823-828

[13] Dinitz JH, Stinson DR. Contemporary Design Theory: A Collection of Surveys. A Wiley-Inter-Science Publication; 1992

[14] La Jolla Covering Repository. Available from: http://www.ccrwest. org/cover.html

Chapter 4

# Electromagnetic Eavesdropping

Ireneusz Kubiak

Abstract

Protection of information against electromagnetic penetration is a huge challenge. Especially this issue applies to computer station that processes protected information and that is a source of electromagnetic disturbances. These disturbances could be correlated with processed graphic information. Therefore, very often, they are called valuable or unintentional emissions. To protect the information, different methods of engineering of electromagnetic compatibility are used, e.g. electromagnetic gaskets, signal and power filters and electromagnetic shielding. The use of these methods causes a special device to become very heavy, and the looks of such device aren't nice. A new universal solution based on safe fonts is proposed. Safe fonts protect processed information against electromagnetic penetration in each case of graphic source of valuable emissions. These fonts protect not only Video Graphics Array (VGA) but also Digital Video Interface (DVI) standards. These fonts are also useful from electromagnetic protection's point of view in the case of the use of laser printers. All analyses are based on images reconstructed from valuable emissions. These emissions are measured in a range of frequencies from 100 MHz to 1.5 GHz. Safe fonts are simple solution that counteract electromagnetic eavesdropping process. They can replace expensive solutions based on shielding, zoning and filtering.

Keywords: electromagnetic eavesdropping, leakage information, protection of information, valuable (sensitive) emission, electromagnetic infiltration process, image and signal processing, data acquisition, identification, recognition

1. Introduction

Protection of information against electromagnetic eavesdropping in modern electronic systems is a big challenge. Such kind of eavesdropping is connected with electromagnetic emissions which are correlated with processed information [1–3]. This problem increases with a higher and higher use of electronic devices for processing and transmitting information. It results from the fact that each electronic device is the source of electromagnetic disturbances in particular unintentional emission of electromagnetic energy transited in surrounding space. Very often signals of unintentional emission could be correlated with processed information. The electromagnetic emissions with distinctive features can arise at any stage of processing of information (e.g. transmission, displaying on screen, printing) which occur in the electric form [1, 4–8]. Video signals are particularly dangerous (**Figure 1**) [9, 10].

Video graphics array (VGA) and digital video interface (DVI) [11, 12] are video standards currently used among other things in nonpublic information systems (other elements—keyboard, screen and main unit—of computer station could be also sources of sensitive emissions [13, 14]). But the graphic lines (VGA, DVI and laser printers) are most susceptible to electromagnetic eavesdropping. These
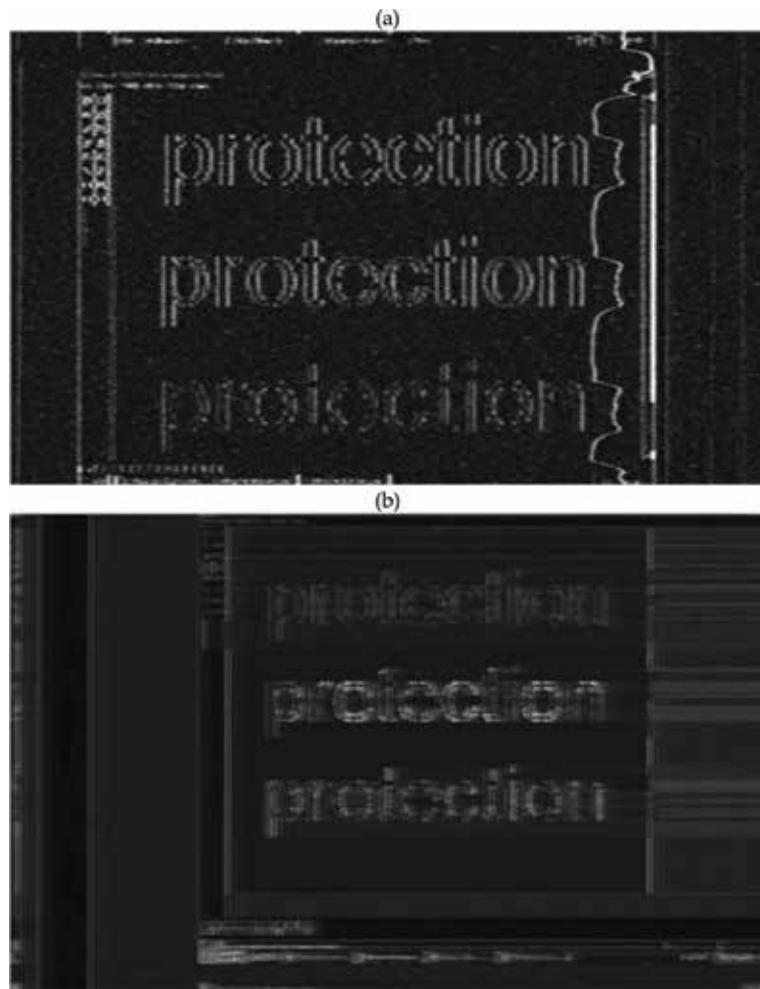
Figure 1.
Examples of reconstructed images for sources of sensitive emissions in the form of (a) VGA standard (receiving frequency f_o = 681 MHz, measure bandwidth BW = 10 MHz) and (b) DVI standard (receiving frequency fo = 852 MHz, measure bandwidth BW = 50 MHz).

standards are the object of the research on solutions that effectively protect processed graphical data. Most frequently, the only solutions used for electromagnetic data protection are those design-related which decrease the level of unwanted emissions at a source. Many people consider which standard is safer.

Another no less dangerous source of electromagnetic emission, formidable from the point of view of the possibility of conducting electromagnetic "watch", are computer laser printers [15, 16]. They translate the electronic form of processed data into graphical form during the printing process. As with every electronic device, printers are sources of electromagnetic emanations [17]. Besides control signals, which carry no information (e.g. directing the operation of stepper motors or heaters), there are other signals (useful signals) that are correlated with the information being processed. Such emissions are called "sensitive" or "valuable" or "compromising" emanations from the point of view of electromagnetic protection of processed information. Processed data may be information displayed on a computer screen or printed (**Figure 2**).

Organizational and technical solutions are the most often used methods for limiting infiltration sensitivity of devices. Technical solutions are limited to changes in
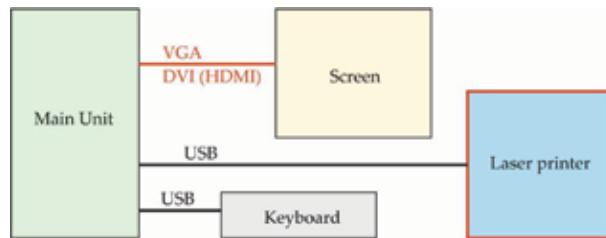
Figure 2.
Sources of sensitive emissions on an example of a personal computer (PC).

the design of devices that typically increase the cost of such devices and sometimes limit their functionality. Therefore, it is desirable to find solutions that avoid these drawbacks and at the same time allow "safe" processing of classified information. An example of an organizational solution might be the establishment of a "control zone" around susceptible devices, relying on distance to attenuate signals below levels that can be received outside the control zone. Both solutions aren't acceptable by users of these devices.

Note that the costs of acquiring a single computer set, referred to as TEMPEST class, are an expense of several of thousand dollars. That is why "software solutions", based on the use of "safe fonts", are mentioned more and more often. As shown by the results of conducted studies, classified information processed with the use of them becomes safe for sources in both the form of video track standard VGA and DVI and video track of laser printers. In the case of laser printers, one technical method that is commonly used in the field of electromagnetic compatibility—both to reduce the amount of electromagnetic interference emitted from the device and the susceptibility of the device to electromagnetic disturbance—is the use of differential-mode signals. This solution protecting printed information was described in [18].

Currently there are new searched methods based on software solutions. Such solutions could change the character of radiation source. The methods could be used to support other solutions or they could be used alone. One of them is safe fonts [19, 20]. Very often such fonts are called TEMPEST fonts. There are three proposed sets of such fonts: Symmetrical Safe font, Asymmetrical Safe font and Simply Safe font (**Figure 3**). These sets of fonts differ in properties of construction of font characters.

Usefulness of these fonts was confirmed from electromagnetic protection's point of view for analogue graphic standard VGA, digital graphic standard DVI and laser printers. The collections of these fonts also are resistant to optical character recognition (OCR).

A side-channel attack (SCA) plays a very important role in the electromagnetic eavesdropping process. The SCA is built from a source of emission, a receiver of emission and space between these two mentioned elements (**Figure 4**).

This type of SCA has the characteristics of a high-pass filter, which is an important property from the protection of information against electromagnetic infiltration process' point of view. The SCA is described by formula.

$$y'(t) = \lim_{\Delta t \to 0} \frac{x(t - \Delta t) - x(t)}{\Delta t}, \qquad (1)$$

where $y(t)$ is the signal (sensitive emission) on the output of SCA and $x(t)$ is the signal on the input of SCA. Very often on the output of receiver, a module $y''(t)$ of the signal $y(t)$ exists:

$$y''(t) = \left| \lim_{\Delta t \to 0} \frac{x(t - \Delta t) - x(t)}{\Delta t} \right|. \qquad (2)$$

In real conditions we cannot forget about noises $s(t)$ and electromagnetic disturbances $d(t)$ which could disturb sensitive emissions $y'''(t)$ measured on the output of the SCA [21]. Then.

$$x^{'(t)} = x(t) + s(t) + d(t), \qquad (3)$$

$$y'''(t) = \left| \lim_{\Delta t \to 0} \frac{x'(t - \Delta t) - x'(t)}{\Delta t} \right|. \qquad (4)$$

On the output of SCA, only vertical and diagonal edges (rising edges and falling edges of pulses of electrical video signals, as on **Figure 1a**) are visible on reconstructed images. There are no visible horizontal edges. It is a very important feature from electromagnetic penetration process' point of view.
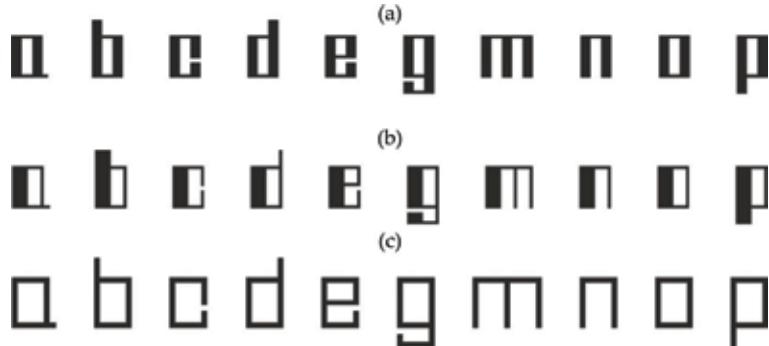
Figure 3.
Examples of characters of safe fonts: (a) symmetrical safe font, (b) asymmetrical safe font and (c) simply safe font.
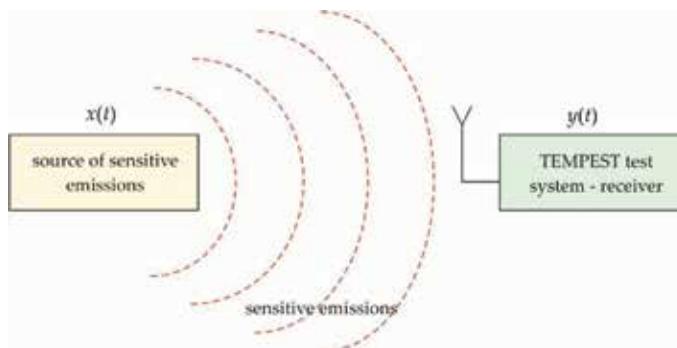
Figure 4.
A scheme of side-channel attack.

2. TEMPEST fonts

2.1 Introduction

An important element of daily processing of text information is the use of computer fonts. Traditional Arial and Times New Roman fonts are the most

popular. Characters of these fonts have decorative elements such as an ear, a bowl, an eye, a serif, a tail, a terminal, a bracket, a loop, etc. The characters are oval, and angles between the individual elements of characters aren't equal to 90° [22, 23]. In addition, the widths of lines building characters are variable. During the processing of text data, each character of font has its representation in the form of electrical signal. This signal is transmitted from a computer to a screen or to a laser printer. In this case, this signal becomes a source of electromagnetic emissions which have characteristics of this signal.

Safe fonts are designed according to safety criteria [24, 25]:

- Lines building the characters intersect at a right angle (each character is built only from vertical and horizontal lines).

- Font characters are devoid of decorative and diagonal elements.

- General contour of characters of safe font has a rectangle shape.

The safe fonts are fully usable. They are resistant to electromagnetic infiltration process as well as OCR process.

The traditional fonts do not meet the mentioned requirements (**Figure 5**).

2.2 Symmetrical safe font

The font characters are devoid of decorative and diagonal elements. The lines building the characters intersect at a right angle. Each character is built from lines about two widths (**Figure 6**). Wider lines are vertical lines of the character; thinner lines are horizontal lines of the character. Simultaneously the right proportions of the line width and the clearance of each character of the font are maintained. It means that the distance between two wider vertical lines is equal to the width of vertical line. The corresponding characters have ascender and descender. There aren't unnecessary decorative elements. This makes that the characters of the font are similar with each other with high values of correlation coefficient between characters.

2.3 Asymmetrical safe font

Similar to the Symmetrical Safe font, this font could be used in printing process and computer techniques. The characters of this font are devoid of decorative and diagonal elements. The lines building the characters intersect at a right angle. Each character is also built from lines about two widths (**Figure 7**). However, the location of the lines in the characters is different than for the Symmetrical Safe font. Wider lines are vertical lines but only as a left part of the character. Thinner lines appear as horizontal lines of the character and as a right element of the character. It means that the width of wider vertical line is the equal sum of distance between vertical lines and width of thinner vertical line. Simultaneously the right proportions of the width of the lines and the clearance of each character of font are maintained. The corresponding characters have ascender and descender.

Figure 5.
Distinctive features of times new Roman font on an example of "a" character.
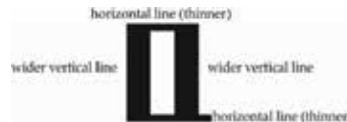
Figure 6.
A construction of characters of symmetrical safe font.

Figure 7.
A construction of characters of asymmetrical safe font.

Figure 8.
A construction of characters of simply safe font.

2.4 Simply safe font

The Simply Safe font is the third set of safe fonts. The characters of this font are devoid of decorative and diagonal elements. The lines building the characters intersect at a right angle. Each character is built from lines about one width (**Figure 8**). This feature distinguishes this font from two others. Simultaneously the right proportions of the width of the lines and the clearance of each character of font are maintained. The corresponding characters have ascender and descender.

3. Sensitive emissions and possibilities of reconstruction of primary information

3.1 Introduction

To assess sensitive emissions from electromagnetic protection of information's point of view, a lot of tests were carried out. The tests were conducted in special conditions without additional unwanted sources of electromagnetic disturbances [26, 27]. Such conditions exist inside an anechoic chamber (**Figure 9**). Such type of chamber is built from metal sheets. Internal walls of the chamber are covered with special hybrid material. These materials (graphite tiles and so-called cones containing graphite compounds) absorb an energy of electromagnetic waves. There is lack of secondary electromagnetic waves.

Sensitive emissions are measured in the range of frequency from about 100 to 1500 MHz. The upper limit of frequency is pointed by digital video standard. For analogue standard the upper limit is equal to about 800 MHz. Of course, it depends on the parameters of image displayed on a screen.

Corresponding tests were carried out to show the effectiveness of the tempest fonts in the protection of information against electromagnetic penetration process. In

Figure 9.
A view of anechoic chamber with hybrid material.

analyses we can use a visual method. It applies to images obtained based on sensitive emissions measured on corresponding frequencies (**Figure 10**, selected frequencies).

Parameters of displayed image:

- Resolution (1024 × 768/60 Hz)

- Size of font characters (22 p.)

- Magnification (600%)

Analyzing the images we have to remember the size of characters and the magnification. For typical parameters, i.e. 12 p. of the size and 200% of the magnification, the recognition process is much harder and even impossible for safe fonts. Very often a coefficient of character error rate (CER)

$$CER = \frac{m + k}{q} = \frac{m + (u - n)}{q},\qquad(5)$$

where u is the number of characters looked for in an analyzed image, m is the number of characters incorrectly recognized, n is the number of characters correctly recognized, k is the number of unrecognized but looked for characters in an analyzed image ($k = u - n$), and q is the number of all characters existing in an analyzed image and is used for precise assessment. The values of this parameter were shown on **Table 1**.

3.2 VGA standard

Signal compatible with the VGA standard has an amplitude ranging from 0 to 0.7 V, and the maximum number of signal levels, when using all 8-bits to encode the brightness of each component of the pixel's color, is equal to 256. This means 16, 777, 216 colors are possible to be displayed with the red, green and blue (RGB) color palette and correspond to the true color (24-bits) graphic card work mode.

Video signal of the VGA standard has a characteristic structure which enables to distinguish unique signatures, which allow its identification, both in the time and the frequency domain.

A sensitive emission was measured among other things on the frequency equal to 417 MHz (**Figure 10a**). The emission allowed to reconstructed primary information [28–30] which was shown on **Figure 11**.
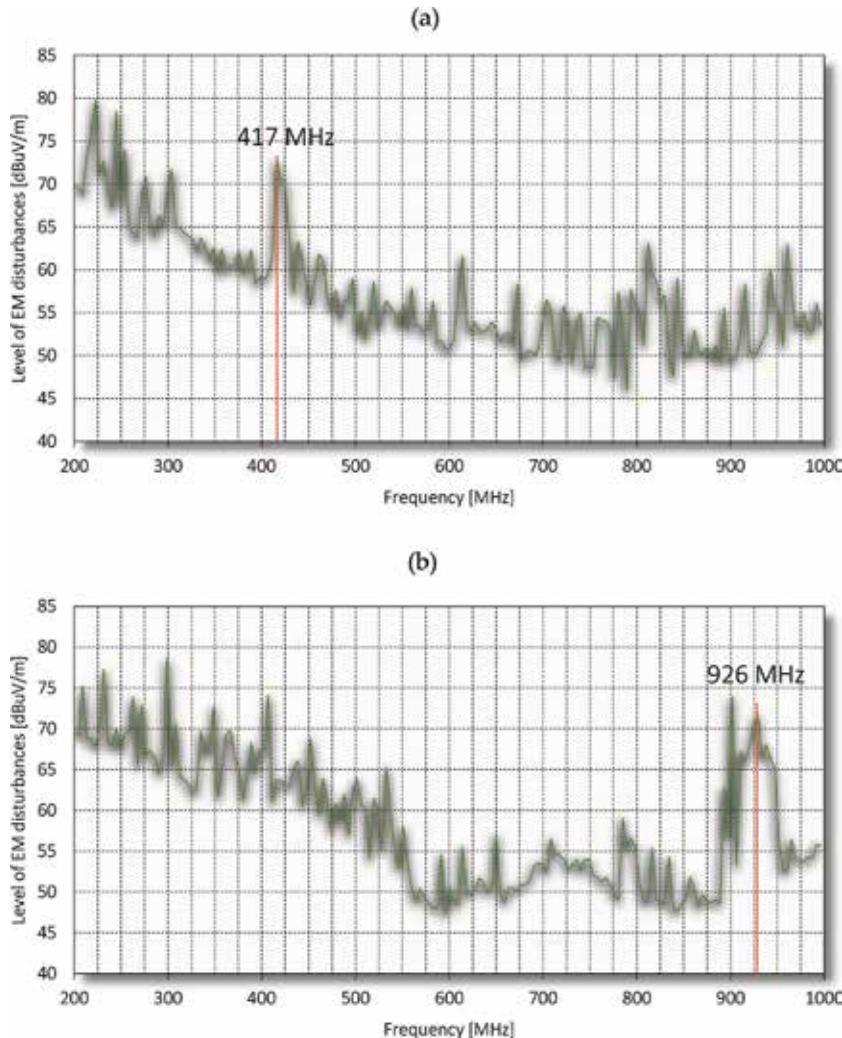
Figure 10.
Levels of electromagnetic emissions measured in frequency domain: (a) VGA standard (b) DVI standard.

3.3 DVI standard

DVI standard specification, developed by the Digital Display Working Group (DDWG), gathering many leading hardware manufacturers, was published in 1999 [11]. In this encoding method, 8-bit RGB data are converted in the graphic card's transmitter into a 10-bit format using transition minimalization and constant component balancing (DC-balanced sequence) techniques. However, that does not mean that transition-minimized differential signalling (TMDS) encoding and extension DVI standard are impervious to electromagnetic eavesdropping. An electric signal in the form of a series of zeros and ones is a source of sensitive emissions. Radiated emission propagates in the space surrounding the source and is subjected to the effects of a high-pass filter (SCA). As a result of the original signal (TMDS) distortions of this kind, the form of sensitive emission signal becomes usable for electromagnetic eavesdropping.

Time series of signals responsible for transmission of information on pixels' color components in DVI standard show that the DVI interface retains the framing principles of signals from the VGA interface. Bit (impulse) series corresponding to individual pixels of the image are transmitted in accordance with the TMDS clock in

Character	Arial font	Times new roman font	Symmetrical safe font	Asymmetrical safe font	Simply safe font
VGA standard					
d	0	0	0	371	1
e	1	6	122	354	48
i	43	17	1	128	115
o	3	9	339	212	174
u	1	2	369	45	107
DVI (HDMI) standard					
d	4	4	25	95	72
e	4	4	140	45	65
i	46	14	42	15	153
o	36	7	60	236	351
u	7	3	104	562	6

Table 1.
Values of character error rate for selected characters of traditional and safe fonts.
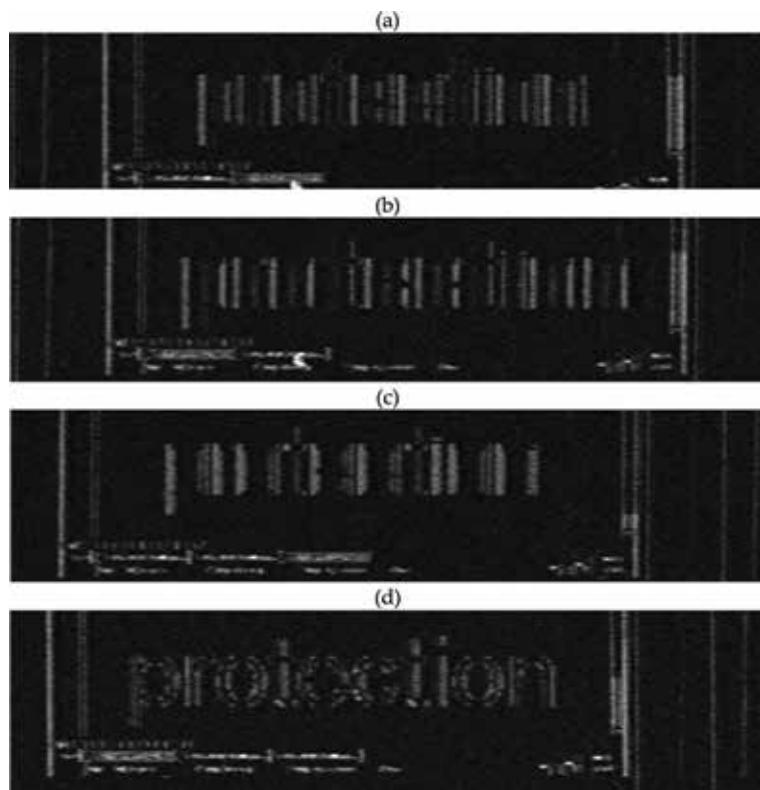
Figure 11.
Fragments of reconstructed images including characters of (a) symmetrical safe font, (b) asymmetrical safe font, (c) simply safe font and (d) Arial font.

strictly defined time periods. Those periods reflect horizontal and vertical synchronization signals of the analogue VGA signal.
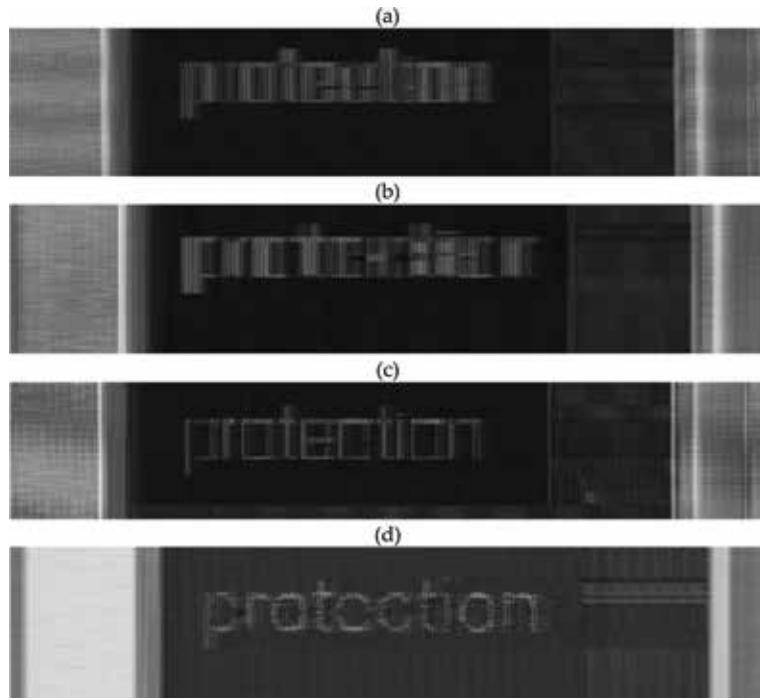
Figure 12.
Fragments of reconstructed images including characters of (a) symmetrical safe font, (b) asymmetrical safe font, (c) simply safe font and (d) Arial font.

A sensitive emission was measured among other things on the frequency of 926 MHz (**Figure 10b**). The emission allowed to reconstructed primary information which was shown in **Figure 12**.

4. Conclusions

In this chapter the possibilities of electromagnetic penetration for sources of sensitive emissions in the form of analogue (VGA) and digital (DVI, HDMI) video standards were described. A solution, which protects processed text information, was shown. The solution is based on three sets of tempest font: Symmetrical Safe font, Asymmetrical Safe font and Simply Safe font.

Each font is free from distinctive features. Characters of fonts are characterized by very high level of similarity. Values of correlation coefficient are more than 0.7, but the level of legibility is acceptable by potential users. The high level of similarity of characters on the input of SCA causes that on the output of SCA the recognition of characters is very difficult. For traditional fonts (e.g. Arial or Times New Roman), the infiltration process is possible and data acquisition is not difficult (**Figure 13**) [31, 32]. Additionally the tempest fonts counteract the penetration process of laser printers. Therefore the new solution is called as a universal solution.

An application of new solution is very easy. Only an installation corresponding set font on a computer is necessary [19]. The solution can replace present solutions basing on shielding (very heavy elements), filtering, grounding and so on.

The safe fonts obtained protection of the Polish Office Pattern in the form of Industrial Design (No. 24487, **Figure 14a**) and Patent (No. 231691, **Figure 14b**) [24, 25].

Figure 13.
Example of reconstructed image for text size 14 p. and magnification 160% written using safe and traditional fonts.

Figure 14.
A certificate of Industrial Design (a) and a Patent (b) for safe fonts.

Conflict of interest

The author declares no conflict of interest of this book chapter.

Author details

Ireneusz Kubiak
Military Communication Institute, Poland

*Address all correspondence to: i.kubiak@wil.waw.pl

IntechOpen

References

[1] Lee HK, Kim JH, Kim SC. Emission security limits for compromising emanations using electromagnetic emanation security channel analysis. IEICE Transactions on Communications. 2013;**E96-B**(10):2639-2649. DOI: 10.1587/transcom.E96.B.2639

[2] Nan Z, Yinghua L, Qiang C, Yiying W. Investigation of unintentional video emanations from a VGA connector in the desktop computers. IEEE Transactions on Electromagnetic Compatibility. 2017;**59**(6):1826-1834

[3] Jinming L, Jiemin Z, Taikang L, Yongmei L. The reconstitution of LCD compromising emanations based on wavelet denoising. In: 12th International Conference on Computer Science and Education (ICCSE); 22-25 August 2017; Houston, USA. pp. 294-297. DOI: 10.1109/ICCSE.2017.8085505

[4] Kubiak I. Laser printer as a source of sensitive emissions. Turkish Journal of Electrical Engineering and Computer Sciences. 2018;**26**(3):1354-1366. DOI: 10.3906/elk-1704-263

[5] Kubiak I. LED printers and safe fonts as an effective protection against the formation of unwanted emission. Turkish Journal of Electrical Engineering and Computer Sciences. 2017;**25**(5):4268-4279. DOI: 10.3906/elk-1701-128

[6] Litao W, Bin Y. Analysis and measurement on the electromagnetic compromising emanations of computer keyboards. In: Seventh International Conference on Computational Intelligence and Security; 3-4 December 2011; Sanya. pp. 640-643. DOI: 10.1109/CIS.2011.146

[7] Tokarev AB, Pitolin VM, Beletskaya SY, Bulgakov AV. Detection of informative components of compromising electromagnetic emanations of computer hardware. International Journal of Computer Technology and Applications. 2016;**9**:09-19

[8] Boitan A, Bărtușică R, Halunga S, Bîndar V. Video signal recovery from the laser printer LCD display, Proceedings Volume 10977, Advanced Topics in Optoelectronics, Microelectronics, and Nanotechnologies IX, 2018. Event: Advanced Topics in Optoelectronics, Microelectronics and Nanotechnologies IX, 2018, Constanta, Romania. DOI: 10.1117/12.2324759

[9] Butnariu V, Trip B, Macovei A, Rosu G, Boitan A, Halunga S. Power line compromising emanations analysis. Annals of the University of Craiova, Electrical Engineering Series. Vol. 48. 2018

[10] Macovei A, Boitan A, Trip B, Halunga S. Detection of electromagnetic emissions transmitted on the power line through electrical conduction. In: International Conference on Applied and Theoretical Electricity (ICATE 2018); DOI: 10.1109/ICATE.2018.8551437

[11] Digital Visual Interface DVI Revision 1.0. Available from: http://www.cs.unc.edu/~stc/FAQs/Video/dvi_spec-V1_0.pdf

[12] HDMI Specification Ver.1.3a. Available from: https://www.hdmi.org/manufacturer/specification.aspx

[13] Jun S, Yongacoglu A, Degang S, Dong W. Computer LCD recognition based on the compromising emanations in cyclic frequency domain. In: IEEE International Symposium on Electromagnetic Compatibility, Ottawa, Canada; 25-29 July 2016; pp. 164-169

[14] Kubiak I. Video signal level (colour intensity) and effectiveness of

electromagnetic infiltration. Bulletin of the Polish Academy of Sciences—Technical Sciences. 2016;**64**:207-2018. DOI: 10.1515/bpasts-2016-0023

[15] Kubiak I. The influence of the structure of useful signal on the efficacy of sensitive emission of laser printers. Measurement. 2018;**119**:63-76. DOI: 10.1016/j.measurement.2018.01.055

[16] Loughry J, Umphress DA. Information leakage from optical emanations. ACM Transactions on Information and System Security. 2002;**5**:262-289

[17] Michael JM. The pentagon worries that spies can see its computer screens, someone could watch what's on your VDT. The Wall Street Journal. 2000;**8**:1-2

[18] Kubiak I, Loughry J. LED Arrays of Laser Printers as Sources of Valuable Emissions for Electromagnetic Penetration Process, arXiv:1808.03007v1 [cs.CR]; 9 August 2018

[19] Kubiak I. TEMPEST font counteracting a non-invasive acquisition of text data. Turkish Journal of Electrical Engineering and Computer Sciences. 2018;**26**(1):582-592. DOI: 10.3906/elk-1704-9

[20] Kubiak I. Influence of the method of colors on levels of electromagnetic emissions from video standards. IEEE Transactions on Electromagnetic Compatibility. 2018:1-9. DOI: 10.1109/TEMC.2018.2881304

[21] Song TL, Jeong YR, Jo HS, Yook JG. Noise-jamming effect as a countermeasure against TEMPEST during high-speed signaling. IEEE Transactions on Electromagnetic Compatibility. 2015;**57**(6):1491-1500

[22] Shainir A, Rappoport A. Extraction of typographic elements from outline

representations of fonts. Computer Graphics Forum. 1996;**15**(3):259-268. DOI: 10.1111/1467-8659.1530259

[23] Nanxuan Z, Ying C, Lau WH. Modeling fonts in context: Font prediction on web designs. Computer Graphics Forum. 2018;**37**(7):385-395. DOI: 10.1111/cgf.13576

[24] Kubiak I. Industrial Design No. 24487, Polish Pattern Office, 10 September 2018

[25] Kubiak I. Pattern No. P.408372 Method for protecting transmission of information, Polish Pattern Office, 3 December 2018

[26] Song TL, Jong-Gwan Y. Study of jamming countermeasure for electromagnetically leaked digital video signals. In: IEEE International Symposium on Electromagnetic Compatibility; 1-4 September 2014; DOI: 10.1109/EMCEurope.2014.6931078

[27] Yuan K, Grassi F, Spadacini G, Pignari SA. Crosstalk-sensitive loops and reconstruction algorithms to eavesdrop digital signals transmitted along differential interconnects, IEEE Transactions on Electromagnetic Compatibility. 2017;**59**(1):256-265

[28] Juric I, Nedeljkovic U, Novakovic D, Pincjer I. Visual experience of noise in digital images. Tehnicki Vjesnik-Technical Gazette. 2016;**23**(5):1463-1467

[29] Lorencs A, Sinica-Sinavskis J. One method of image processing and its numerical analysis. Elektronika IR Elektrotechnika. 2010;**103**(7):25-30

[30] Manson J, Schaefer S. Wavelet rasterization. Computer Graphics Forum. 2011;**30**(2):395-404. DOI: 10.1111/j.1467-8659.2011.01887.x

[31] Zoric A, Perisic D, Obradovic S, Spalevic P. Virtual multisensors data

acquisition and analysis system design. Elektronika IR Elektrotechnika. 2011;**116**(10):49-54. DOI: 10.5755/j01. eee.116.10.879

[32] Kuusik A, Reilent E, Lõõbas I, Luberg A. Data acquisition software architecture for patient monitoring devices. Elektronika IR Elektrotechnika. 2010;**105**(9):97-100

Geo Location of Mobile Device

Bashar M. Nema and Ali Nafaa Jaafar

Abstract

The proliferation of cellular network enabled users through various positioning tools to track locations; location information is being continuously captured from mobile phones, where a prototype is created that enables to detect locations based on using the two invariant models for Global Systems for Mobile (GSM) and Universal Mobile Telecommunications System (UMTS). The smartphone application on an Android platform applies the location-sensing run as a background process, and the localization method is based on cell phones. The proposed application is associated with remote server and used to track a smartphone without permissions and Internet. Mobile stores data location information in the database (SQLite) and then transfers it into location API to obtain locations' result implemented in Google Maps. Track a smartphone with fixed identifiers mostly SSN (SIM (Subscriber Identity Module) Serial Number) and IMEI (International Mobile Equipment Identity) derived from an identifying string unique to the user's device. The location result is moderately correct according to the cellular networks GSM and UMTS, which are used for obtaining location information.

Keywords: tracking, GSM/UMTS, Web services, MySQL, Mobile device

1. Introduction

Engineers and researchers have studied the cellular wireless systems offering reliable mobile location estimates for the past few years because of its temporal/spatial nature and rich context [1]. The mobile device's location has become an important topic in wireless communication despite its low cost and transmission over the control channel owning short response time. The GSM/UMTS can be used to obtain location information without additional external hardware when there is cellular coverage [2]. Location-based services (LBS) for mobile devices have grown rapidly over the past few years and forecasts show similar growth in the future. As the need for tracking mobiles in our daily life is increasing, it became tracking thousands of users periodically; tracking mobiles by cellular networks does not need any hardware change within the core network. Tracking a smartphone is possible by exploiting a permanent mobile data link [3].

The mobile device used the location information consisting of Cell Identifier (cell ID), Location Area Code (LAC), Mobile Network Code (MNC), and Mobile Country Code (MCC). The location information is stored inside the SQLite database in mobile device periodically [4].

The contributions in this chapter can be summarized as follows: we utilize location information for a mobile device that is always available in the cellular network. We create an application of an efficient, easy-to-use, and inexpensive mobile device tracking system; the application provides the better balance between

location accuracy and battery power. The location information collected during the mobile device roaming between the base stations connects them and the location information enables us to track the mobile device to see the historical location; the application can be used in any place in world covered by cellular network GSM/UMTS, compared to the information of location with the location APIs to provide the mobile device location. This work offers proactive services of assistance that consequently tell their users when they enter or leave the limits of pre-characterized focal points.

2. Structural GSM/UMTS architecture

The GSM network structure is divided into a base station subsystem and core network, which is shown in the **Figure 1**. The Base Transceiver Station (BTS) is a main communication area with mobile phones. BTS is linked to the Base Station Controller (BSC), which controls the small set of BTSs. These are linked using an interface known as the (A-bis) interface. The BSC handles the radio channel setup and frequency hopping control and links with a Mobile Switching Center (MSC) via the interface termed the (A interface). The core of a network subsystem is an Authentication Center (AuC) that provides authentication, allowing users to access the network. MSC via interfaces with the Visitor Location Register (VLR) and Home Location Register (HLR) provides location information for a network [5].

The connections originate in the fixed network (Public Switched Telephone Network (PSTN) and Integrated Services Digital Network (ISDN)). The connections have to pass through a dedicated Gateway Mobile Switching Center (GMSC). SIM card Serial Number (SSN) is stored on the SIM card. SIM is inserted to the mobile for providing information around the subscriber. IMEI consists of 15 digit number to identify a mobile phone. IMEI is checked when the mobile phone is
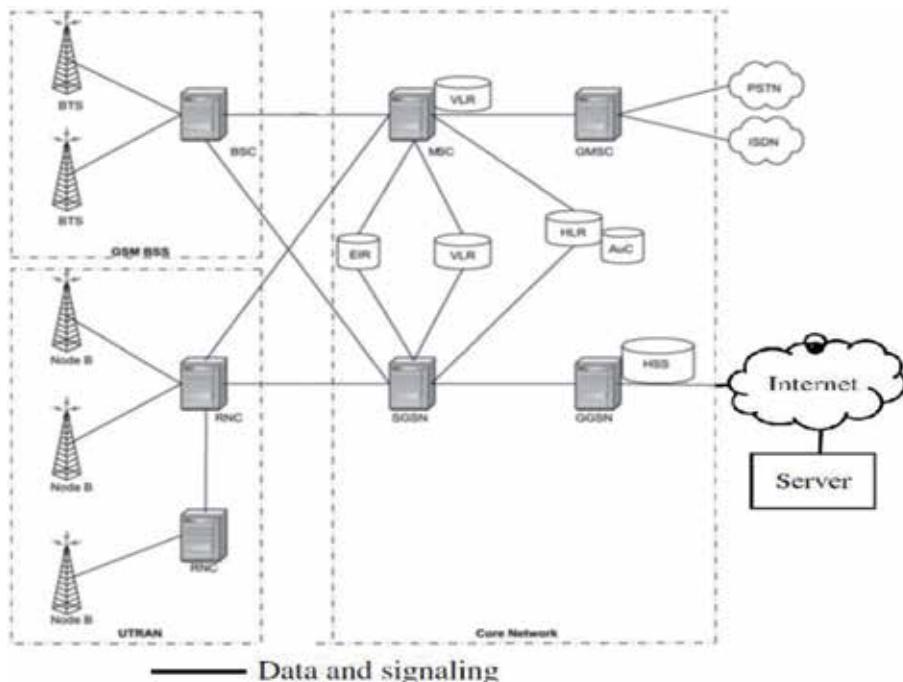
Figure 1.
The architecture GSM/UMTS network [5].

trying to access the network and compared with an Equipment Identity Register (EIR) database. EIR registers equipment data rather than subscriber data [6].

UMTS network structure is divided into the UMTS Terrestrial Radio Access Network (UTRAN) and core network. UTRAN includes two nodes (the NodeB and Radio Network Controller (RNC)). Each RNC controls one or more NodeBs and is responsible for the control of radio resource parameters of cells managed by those NodeBs [7].

The Home Subscriber Server (HSS) is a database that communicates directly with a Gateway GPRS Support Node (GGSN) and Service GPRS Support Node (SGSN). SGSN, HSS, and GGSN exchange signaling messages only. Every GGSN can serve multiple SGSNs and every SGSN can serve multiple RNCs. RNC provides an air interface (Iub) between a mobile terminal and a core network, and an air interface (Iur) between RNCs.

There are two fundamental differences among GSM and UMTS systems. GSM networks use a methodology of more than one database communication (HLR and VLRs). The neighboring RNCs can ability to communicate with one another. In GSM systems, there is no immediate correspondence between neighboring VLRs and the correspondence is occurring just through the HLR. UMTS systems utilize a methodology with just a single database (HSS) to record the movements of the portable terminals [5].

3. GSM/UMTS-based localization

Cell-ID (ID) in combination with the MCC, the LAC, and the MNC is the unique identifier of the BTS. The Cell-ID of the BTS has the association with certain Mobile Station (MS), which is known by the mobile device and can be utilized to appraise the position of the mobile device [8]. GSM/UMTS service zone is separated into Location Areas (LAs), where every LA incorporates at least one radio cells. Every LA and radio cell has a unique identifier named Cell-ID and LAC [9].

BTS covers the set of cells, each of them identified by the unique Cell-ID (C1, C2, and C3 as shown in **Figure 2**). MS persistently chooses a cell and trades information and signaling traffic with the relating BTS. The cells are assembled into groups, every one of them distinguished by a Location Area Identifier (LAI). So as to maintain a strategic to avoid excessive signaling traffic, as long as the MS is in idle mode, the system knows just the LAI. The network gets mindful of the Cell-ID just when the MS switches into a dedicated mode, in particular when the channel is utilized to
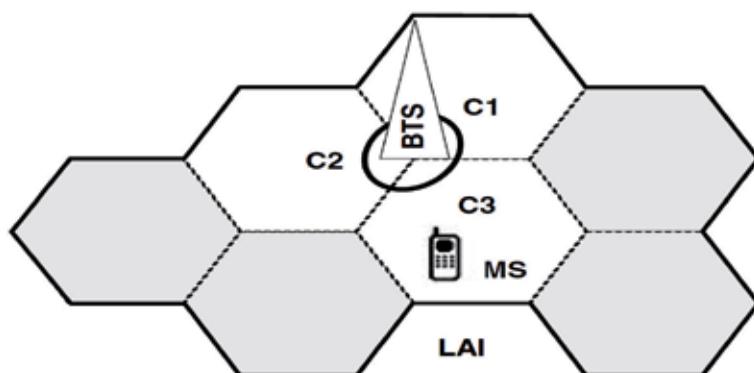
Figure 2.
A simplified view of the cellular network and the mobile station is within the cell C3 [10].

Cell type	Cell dimension (meter)
Nanocell	1–10
Picocell	10–100
Microcell	100–1000
Small macrocell	1000–3000
Large macrocell	3000–30,000

Table 1.
The size range of cell type [10].

really establish a call. MS always knows the Cell-ID of the cell it is in. The size of a cell may range some meters; accuracy depends on cell size (illustrate in **Table 1**) [10].

4. Android platform

Android is the platform introduced in 2007 for devices such as smartphones or tablets developed by Google, which is a Linux-based operating system and an open source code designed for touch screen mobile device. The applications were developed in Java language, which allows the software to be freely modified using the Android Studio software development kit (SDK). The SDK contains a comprehensive set of software libraries supported by integrated development environment (IDE), the Android Studio (Android 7.0) to get the cell ID, LAC, MMC, MNC, IMEI, and SSN of an Android mobile, and the database (SQLite) used to store this cell ID, LAC, MMC, and MNC values into periodical processes that are running as background process by applying Android service; Google Maps get from Web service the latitude and longitude in the format of a Java Script Object Notation (JSON) file, the Android mobile support from both Google Maps and third-party developers (remotely connect to a MySQL database) [11].

5. Web service

A Web service is the software designed to support communication between mobile application and remote server and location API over a network. The Web service designed to provide compute location information results in obtaining specific geographic regions (longitude and latitude) after connected with location API according to request by users; the Web service uses an online portal developed in the PHP which is an open source general; the Web service sends and receives data with insert/delete, and the Web service performs the tasks and generates output in the JSON format [12]:

[{"gsmlatitude":"33.24567","gsmlongitude":"44.362478","datetime":"2017/12/22 17:17:26"},
{"gsmlatitude":"33.245621","gsmlongitude":"44.362425","datetime":"2017/12/22 17:22:12"},
{"gsmlatitude":"33.245621","gsmlongitude":"44.362425","datetime":"2017/12/22 17:28:36"},
{"gsmlatitude":"33.249792","gsmlongitude":"44.358406","datetime":"2017/12/22 17:33:19"},

{"gsmlatitude":"33.249792","gsmlongitude":"44.358406","datetime":"2017/12/22 17:38:55"}]

Representational State Transfer (REST) is a structural style for creating Web services and exploits the advances and conventions of the World Wide Web, the relationship between the mobile and remote server by volley technique method proposed by Google 2013. The plan utilizes POST and JSON configurations to move information [13].

6. Tracking system

In order to implement a mobile tracking system with a GSM/UMTS network of detection of mobile location for places visited through your phone, the mobile device tracker connects with the tower BTS and stores location information (cell id, LAC, MCC, MNC) continuously for each specific period (5 minutes) in
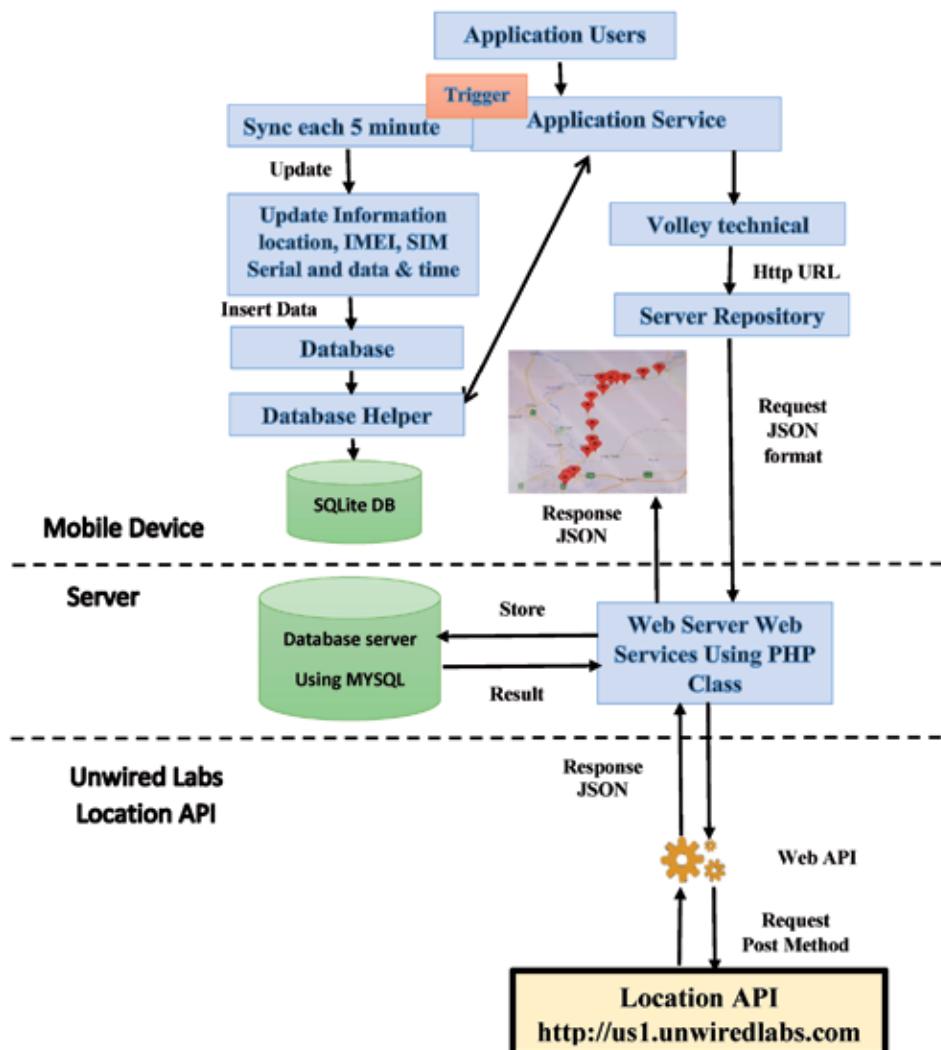

Figure 3.
Proposed tracking system.

SQLite Database. All these processes run as background by applying Android service which is without a suspect by the mobile user. When mobile connected with internet can retrieve location information from SQLite Database of the mobile device, the location information transfer into the remote server during synchronizing between SQLite and MySQL. The server uses PHP and MySQL and can get all the necessary data to locate the phone. The PHP file will compute both latitude and longitude for location information that is stored in MySQL during which connections are created with location API to obtain the locations visited by the mobile device and the results are stored in MySQL. The data (latitude and longitude) is then transferred to the mobile device in the form of the JSON format and Google Maps is used that will plot the locations. **Figure 3** shows the tracking system.

In this section, we propose a tracking system algorithm to develop models of the relationship between the mobile application and the server site. The proposed algorithm can be implemented in six steps, which are described in Algorithm (1):

Algorithm (1): Description of the tracking system:

Step 1: The mobile device can read the location information (cell id, LAC, MCC, MNC) using cellular network (GSM/UMTS) every 5 minutes. The location information, IMEI, SSN, and current date&time were stored in the SQLite database every 5 minutes. Note IMEI, SSN, and current date&time can be obtained from the mobile device.

Step 2: Always have 100 records inside the SQLite database to maintain the storage space of the mobile device. In case of addition, the first field is deleted to guarantee nonexceeded 100 fields.

Step 3: Transfer location information, IMEI, SSN, and date&time from SQLite into the remote database MySQL using Volley Technologies and Web services, the server has received (IMEI, SSN, cell ID, LAC, MNC, MCC) via POST.

Step 4: The Web services provided communication between MySQL and the location API; the location information is then transferred into a location API for obtaining latitude and longitude of each record; the link location API uses URL= http://us1.unwiredlabs.com/process.php to obtain the latitude and longitude.

Step 5: The results store latitude and longitude received from the location API in the MySQL database and create a JSON file.

Step 6: Through the mobile application, the data location is queried from the external database MySQL and received in the form of the JSON file format; the results use a JSON parser to display the location on Google Maps.

The tracking system can be separated into two parts that are mobile application and server site.

• Mobile application

The Android mobile application obtains IMEI, SSN, and current location information (cell id, LAC, MNC, MCC) in the four parameters; the four parameters are considered the basic and latitude and longitude can be found.

These application processes are run in a background service using the Android service. This allows it to run operations without affecting your user so that it can continue to update current location information of the four parameters at specific intervals and periodically for a particular time to be designated by request service while device users move around towers of the mobile phone. After obtaining four parameters it stored in the SQLite database of the internal storage with the IMEI number and the SIM card serial number and the current date and time, this continues to store inside SQLite even if the mobile becomes the screen off and without the need to connect the mobile phone to the Internet, the **Figure 4** shows sequence diagram the mobile tracking.
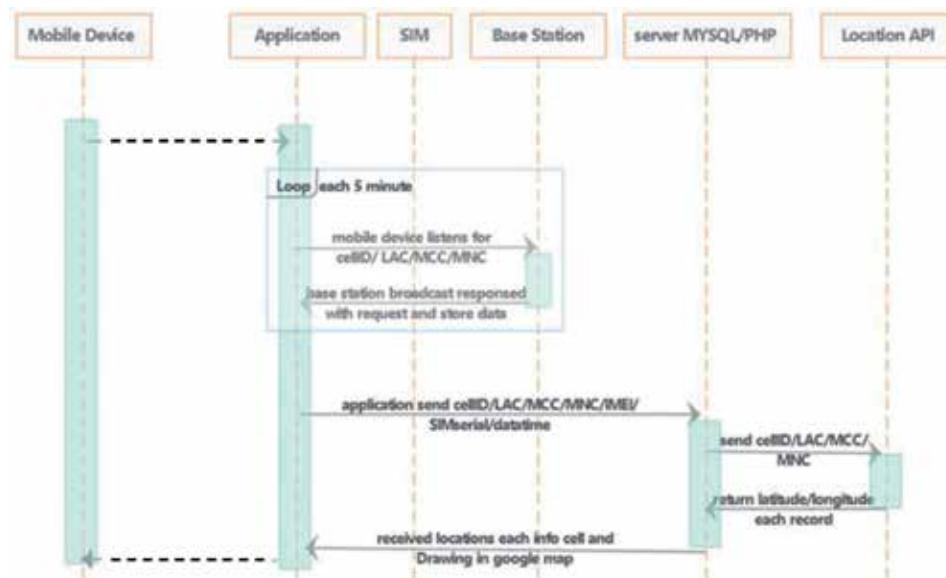
Figure 4.
A sequence diagram of mobile tracking using GSM/UMTS.
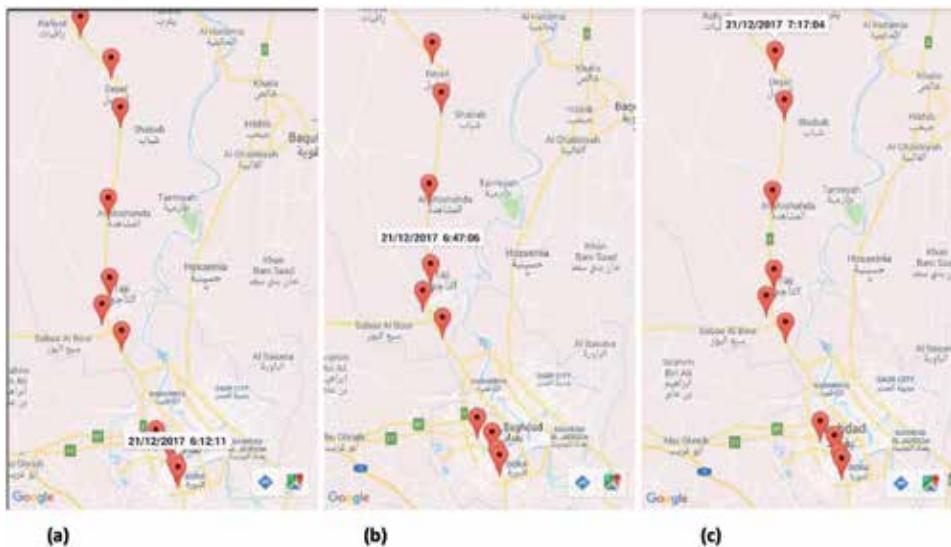
The SQLite database can store a maximum of 100 records which does not affect the storage capacity of the mobile device with the increasing data stored; in case if the record number inside SQLite database exceeds 100 records, the application deletes the first record and adds the current record at the end of the table. After connecting to the Internet, the mobile device can synchronize data between SQLite and MySQL using the JSON format as intermediate.

• Server site

The server side includes a scripting language to be embedded into a PHP source document and a MySQL that is uses the open source relational database management system. After transferring four parameters to MySQL, create a connection between database MySQL and location API site http://us1.unwiredlabs.com so as to retrieve latitude/longitude from four parameters in all records and the output is stored in MySQL database. All records in a table have a specific IMEI number, this send the data to the Android mobile application in JSON format.

7. Results

We have successfully implemented an Android application of a mobile tracking system by using a GSM/UMTS network; the application was run on December 21, 2017, for the distance from Baghdad to Samarra, the location information was recorded every 5 minutes, that can find the nearest position visited, the mobile was not connected with Internet. Some results were shown in **Figure 5**. The results we see after calculating latitude and longitude depending on the location information of the GSM/UMTS network are described in **Table 2**.

Figure 5.
Display result in Google Maps: (a) when the first marker is clicked, data&time are displayed; (b) with middle marker, title is displayed; (c) with the last marker, the title marker data and time is displayed.

The application updates a location information that can also cause heavy battery consumption; this can be lowered to achieve better battery efficiency. The location information updates every 5 minutes the interval at which the current location is updated. **Table 3** illustrates the different intervals that can affect the battery life. The work creates new opportunities to perform tracking on remote servers, using a mobile device only to capture the location information and display the output results. The Volley Technologies provided the good communication infrastructure to transfer a data to the remote server and vice versa. When is receiving 25 requests for transferring the data between the mobile device and the remote server, compare a sync task and volley technology, a sync discussion 13.957 milliseconds and the volley discussion 4.275 milliseconds. Some tracking applications depended on GPS; there exist situations where GPS is not available, and the solution is to combine the GPS measured values with the measured values from the GSM/UMTS. This work provides a service that automatically tells their users when they enter or leave a boundary of predefined points of interest such as application Tammini from Zain.

8. Conclusion

The application showed a low-cost tracking system using a GSM/UMTS network, which is suitable for the works all over the world with the combination of the Android mobile phone and Web services. The overhead is much lower compared to average battery consumption, and the application successfully tracks mobile across a single sensing. We demonstrated tracking using a cheap hardware with open source projects and showed mapping techniques with cell tower databases to take advantage of tracking a mobile device. Finally, we proposed an improvement in location accuracy by using GSM/UMTS tracking that could be implemented without connecting to API location, thus using external geographical information and achieving the best performance for mobile dynamic location with Kalman filters.

id	MCC	MNC	LAC	Cell ID	IMEI	Sim serial	Date&time	Latitude	Longitude
855	418	05	14003	19945606	35943505891918	8996405440003317062	21/12/2017 6:12:11	33.245665	44.362809
856	418	05	14003	19945609	35943505891918	8996405440003317062	21/12/2017 6:17:04	33.245144	44.362708
857	418	05	14003	19949440	35943505891918	8996405440003317062	21/12/2017 6:22:47	33.254439	44.358063
858	418	05	14003	19943755	35943505891918	8996405440003317062	21/12/2017 6:27:52	33.279606	44.349513
859	418	05	14002	19946338	35943505891918	8996405440003317062	21/12/2017 6:32:37	33.30084	44.323802
860	418	05	14008	19807215	35943505891918	8996405440003317062	21/12/2017 6:37:24	33.449417	44.261456
861	418	05	14008	19807215	35943505891918	8996405440003317062	21/12/2017 6:42:22	33.44942	44.26146
862	418	05	14008	19813445	35943505891918	8996405440003317062	21/12/2017 6:47:06	33.489079	44.226421
863	418	05	14008	19820716	35943505891918	8996405440003317062	21/12/2017 6:52:41	33.52829	44.240307
864	418	05	7096	50555	35943505891918	8996405440003317062	21/12/2017 6:57:28	33.647154	44.237929
865	418	05	7096	53415	35943505891918	8996405440003317062	21/12/2017 7:02:37	33.782224	44.259622
866	418	05	7096	53415	35943505891918	8996405440003317062	21/12/2017 7:07:24	33.782224	44.259622
867	418	05	7096	53415	35943505891918	8996405440003317062	21/12/2017 7:12:11	33.782224	44.259622
868	418	05	7096	51317	35943505891918	8996405440003317062	21/12/2017 7:17:04	33.855946	44.242762

Table 2.
Computing latitude and longitude from basis parameters.

Time interval	Details
Every 5 seconds	This provides heavy consumption of battery power
Every 1 minutes	This is the default setting, which provides a better battery power
Every 30 minutes	This provides better battery efficiency

Table 3.
Computing frequency update intervals.

Author details

Bashar M. Nema[1*] and Ali Nafaa Jaafar[2]

1 College of Science, Mustansiriyah University, Baghdad, Iraq

2 Electrical Engineering Technical College, Middle Technical University, Baghdad, Iraq

*Address all correspondence to: bashar_sh77@uomustansiriyah.edu.iq

IntechOpen

References

[1] Assimakis N, Adam M. Global systems for mobile position tracking using Kalman and Lainiotis filters. The Scientific World Journal. 2014;**2014**:130512

[2] Liu H, Zhang Y, Su X, Li X, Xu N. Mobile localization based on received signal strength and Pearson's correlation coefficient. International Journal of Distributed Sensor Networks. 2015;**11**(8):1-10

[3] Ficek M, Pop T, Kencl L. Active tracking in mobile networks: An in-depth view. Computer Networks. 2013;**57**(9):1936-1954

[4] Roberto R, Lima JP, Teichrieb V. Tracking for mobile devices: A systematic mapping study. Computers and Graphics. 2016;**56**:20-30

[5] Vergados DD, Panoutsakopoulos A, Douligeris C. Group registration with distributed databases for location tracking in 3G wireless networks. Computer Networks. 2008;**52**:1521-1544

[6] Bettstetter C, Vogel H-J, Eberspacher J. GSM phase 2+ general packet radio service GPRS: Architecture, protocols, and air interface. IEEE Communications Surveys and Tutorials Surveys. 1999;**2**

[7] Lingyang S, Shen J. Evolved Cellular Network Planning and Optimization for UMTS and LTE. UK: CRC Press/Taylor & Francis Group; 2011

[8] Andreas W, Shangbo W, Guido Bruck H, Peter J. Traffic congestion estimation service exploiting mobile assisted positioning schemes in GSM networks. Procedia Earth and Planetary Science. 2009;**1**:1385-1392

[9] Ntantogian C, Valtas G, Kapetanakis N, Lalagiannis F, Karopoulos G, Xenakis C. Attacking GSM Networks as a Script Kiddie Using Commodity Hardware and Software. Switzerland: Springer International Publishing AG; 2015. pp. 73-86

[10] Trevisani E, Vitaletti A. Cell-ID location technique, limits and benefits: An experimental study. In: Sixth IEEE Workshop on Mobile Computing Systems and Applications. Windermere, Cumbria, UK. 2004;**1**:51-60

[11] Sukapha S. An implementation of location-based service system with cell identifier for detecting lost mobile. Procedia Computer Science. 2011;**3**:949-953

[12] Mantas VM, Liu Z, Pereira AJSC. A web service and android application for the distribution of rainfall estimates and earth observation data. Computers and Geosciences. 2015;**77**:66-76

[13] Arroqui M, Mateos C, Machado C, Zunino A. RESTful web services improve the efficiency of data transfer of a whole-farm simulator accessed by android smartphones. Computers and Electronics in Agriculture. 2012;**87**:14-18

Section 3

Wireless Communication

Chapter 6

Cognitive Radio-Modulation and Demodulation

Madhuri Gummineni and Trinatha Rao Polipalli

Abstract

The reconfigurability in Cognitive Radio (CR) facilitates to dynamically change its parameters for the efficient spectrum utilization. The motivation behind the study of cognitive radio is that the number of different radio signals can be handled without using extra circuitry, i.e., reusing identical hardware with the change in the software will reduce time to market, development cost, and upgrade infrastructure. Software Defined Radio (SDR) is an enabling technology for Cognitive Radio (CR); therefore, it emphasizes on SDR unique features, characteristics, and basics concepts that are required to understand operation of SDR. SDR allows service providers to upgrade infrastructure without unreasonable cost. Modulation techniques play a vital role in any communication systems such as cable modems, DSL modems, CDMA, 4G, Wi-Fi, and WIMAX; thus, it emphasizes on implementation of modulation techniques using SDR Generic hardware, which is operated by Open Source software called GNU Radio. Implementation of various analog and digital modulation techniques using the GNU Radio provides a way for developing advanced wireless communication system. GNU Radio software is a highly flexible signal processing platform, which makes it easy and reduces time to implement different modulation techniques with appropriate script.

Keywords: spectrum scarcity, spectrum management, cognitive radio, software defined radio (SDR), modulation

1. Introduction to cognitive radio

In advanced wireless communication systems, an intelligence of cognitive radio helps to become aware of the environment and internal state and can adapt itself for better performance. Cognitive Radio utilizes Software Defined Radio, Adaptive Radio, and other technologies to automatically adjust its behavior or operations to achieve desired objectives. A Cognitive Radio may adjust all the transmission characteristics from frequency to power level. Software defined radio (SDR) and artificial intelligence (AI) technology enables the field of cognitive radio. A Cognitive Radio is an extension of modern SDR [1]. There is need for cognitive radio in present 21st century due to following reasons.

   i. Generations and its impact on wireless traffic

  ii. Spectrum scarcity

  iii. Spectrum management

Generations	Standards	Technology-data rate	Applications
1G	AMPS, TACS	Analog–2.4 kbps	Only voice
2G (802.11a)	GSM, CDMA, EDBE, GPRS	Digital–64 kbps	SMS and voice
3G (802.11 g)	WCDMA, UTMS	Digital–144 kbps to 2Mbps	WAP, SMS and voice
4G (802.16e)	SEAMLESS MULTIMODE, LTE-A	Digital 802.16 (WiMax)- up to 1 Gbps	Online gaming, TV, Internet, video calls, WAP, SMS, and voice
5G (802.16 m)	Massive MIMO	WWWW–higher than 1 Gbps	High speed applications, mobile TV and wearable devices

Table 1.
Generations and its improvements in applications.

i. **Generations and its impact on wireless Traffic:** Every generation have some standards, different capacities, new techniques and new features which differentiate it from the previous one. The complexity of functions, components and design rules of these architectures continues to increase with each generation, as shown in **Table 1**.

Initially wireless communication network was analog and used for voice calls only. Later digital technology supports text messaging, internet, whereas Fifth Generation Mobile technology is going to be a new revolution in mobile market which has changed the means to use cell phones within very high bandwidth.

These Applications demands high data rate, and create traffic congestion, that intern leads to create requirement for more complex communication standards and modulation techniques. It cannot be achieved with analog platform because analog have inherent fabrication errors. In particular, future seamless multi- mode networks will require radio terminals and base stations with agile, RF bands, channel access modes, data rates, power and functionality.

ii. **Spectrum Scarcity:** The Rules and channel allocation are generally made by FCC. Because of heavy traffic and strict rules mobile users are allowed to use certain frequency only, which get crowded. Mobile frequencies are getting crowded with various applications like Music, Video on the Internet which intern requires more bandwidth.

iii. **Need for Advanced Spectrum Management:** Spectrum is regulated by governmental agencies. Spectrum access is granted by the regulatory agencies to the primary users or licensed to them on a long term basis. Due to the allocation, resources are wasted because large-frequency regions are used very sporadically. The aim of spectrum management [2] is to assign suitable channel/frequency to the mobile users based on the required condition being aware of geographically bounded region, that may be, a personal, local, regional, or global cell. Hence, there is a need for intelligent device in order to use the whole radio spectrum in the most efficient way.

2. Benefits of cognitive radio

The following are the benefits of Cognitive Radio:

i. The base band section of a communication system can be implemented by software which is a unique platform provided by cognitive radio. The theoretical concepts can be verified through experimental observations which motivated to opt Software Defined Radio.

ii. Optimum use of spectrum, organize interoperability, reconfigure networks to meet current needs.

iii. A software radio terminal, for example, could operate in a GSM network, an AMPS network, and a future satellite mobile network. Software radios have emerged to increase quality of service through such agility. The SDR architecture overcomes all the tradeoffs of hardware radio and provides opportunity to customize the performance of the radio.

CR being the central unit, all the users can connect and communicate at the same time. CR can be compatible with any radio. This combined with simple radio embedded in any object would allow interaction with any physical object (**Figure 1**).

iv. Signal transmission and reception between Cognitive radio depends mainly on channel and power, and they will be dynamically changing for improving spectrum efficiency and This process is defined as dynamic spectrum management. The steps for dynamic spectrum Management are

   a. Identify the spectrum hole.

   b. Ensure that spectrum not being used by any users.

   c. Before establish the communication between Cognitive Radios find the common spectrum hole.

   d. During communication if it detects any user it should stop transmission and change to another spectrum hole.

Major challenges of DSM are the parameters of Radio which been changing due to environment and optimal channel assignment.
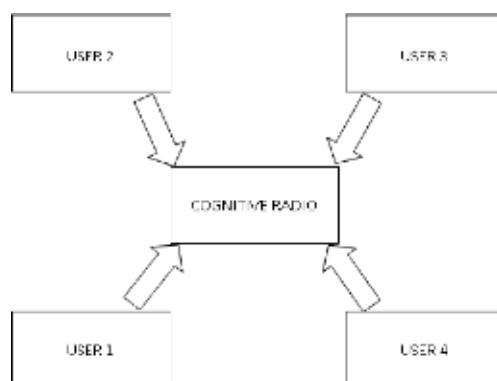
Figure 1.
Cognitive radio as central control unit.

3. Software defined radio (SDR)

The purpose of Radio is to transmit or receive electromagnetic radiation, whereas software facilitates modifiable instructions executed by a programmable processing device [3]. Software defined refers to the use of software for processing within the radio system or device to implement operating functions, in other words a large portion of functionality is implemented through programmable signal processing devices.

Software Defined Radio is a reconfigurable radios because physical layer behavior is mostly altered through software as well modulation waveforms and all wireless communication is controlled by the software. Therefore, this approach reduces the content of RF and other analog components of traditional radios and emphasizes on DSP to enhance overall receiver flexibility (**Figure 2**).

SDR have transformed electronic devices from transistor to software. SDR functionality is partitioned in to software and Hardware domain. The Software domain supports all Signal processing Tasks, whereas IQ samples are handled by Hardware. SDR requires a general purpose RF Front end, which consists of Pre-Selector/Power Amplifier, Receive/Transmit Chain, and Synthesizer.

Pre-Selector/Power Amplifier is required to select subset of spectrum. Receive/Transmit Chain is required to Process the radio signal in analog domain. Synthesizer is required to generate local oscillations for processing.

3.1 SDR overview

Software-defined radio is an integrated platform of software and hardware that enable to develop application based on the requirement in short span of time.

Hardware: HackRF One, Pico Zed SDR, USRP-B200, USRP-B210, BladeRF, UmTRX, Matchstiq, RTL2832U-RealtekSDR dongle,

Software: GNU Radio, Gr-OsmoSDR, SDR#, HDSDR, Gqrx.

Front-End Issues of SDR:

1. HackRF One (Half-duplex)—supported by GNU Radio.

2. Pico Zed SDR -Supported by Matlab and Simulink.

3. USRP-B series (Full-duplex) -supported by GNU Radio.
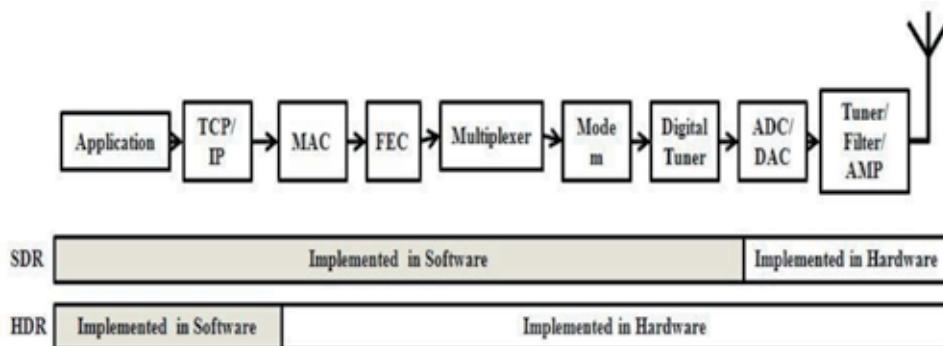
Figure 2.
Depicts the Software Defined Radio and Hardware Defined Radio.

4. The BladeRF 2.0 micro features support: GNU Radio, Gr-OsmoSDR, SDR Sharp-BladeRF, MATLAB, and Simulink.

The hardware aspects of a platform consist of the Radio-Frequency (RF) elements, some baseband signal processing and communications link to the software based signal processing element—perhaps a DSP, FPGA, or a general purpose processor (GPP) (**Figure 3**).

According to functions performed by an SDR, it can be called to be:

i. Multiband system: Multiband system reuse the locally unused spectrum and increase the total system capacity [4, 5]. For implementing multiband system CR will listen to the channel and, if the channel is free it will transmit. One of the application of this system is the transmission area is expanded by using multi-hop networks with multi-band channel. (e.g., GSM 900, GSM 1800, GSM 1900),

ii. A Multi standard system: support more than one standard family (e.g., UTRA-FDD, UTRA-TDD for UMTS) or across different networks (e.g., DECT, GSM, UMTS, WLAN).

iii. Multiservice system: By using Algorithms such as Greedy, Genetic, Multivariable, and Localized Algorithm [6–10] etc. for multi service channel allocation, t allows to provide services to video and data service simultaneously for frequency/time slotted cognitive radio (CR) networks.

iv. A Multichannel system improves the throughput [11], and it can be implemented using CR, by detecting the multiple idle channels, and by suing multichannel access modes such as OFDM and MC-CDMA, thus it covers the multiple channels during the absence of PU.

3.2 Functions of SDR transmitter and receiver

Software radio performs the above functions at the transmitter side and is shown in figure. It identify the available transmission channels [8, 12], determines the suitable channel modulation, examine the propagation path, and steers it's
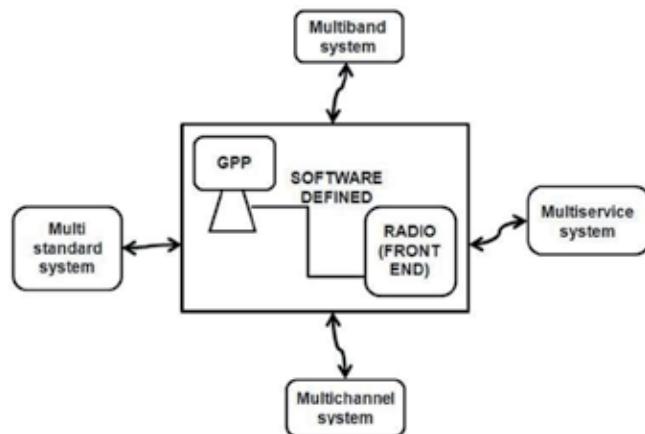
Figure 3.
SDR overview.

transmit beam in the right direction. And also, it determines the appropriate power level, and then transmits in an advanced application.

The following functions are performed at the receive side of SDR. It detects and compares the energy distribution in the channel and in adjacent channels, the mode of the incoming transmission is recognized, interference is nullified adaptively by the receiver. For multipath signal it estimates the dynamic properties for recovering the signal effectively, combines desired-signal from multipath coherently, adaptively equalizes, decodes the channel modulation, and then corrects residual errors via forward error control (FEC),it also decode to receive a signal with lowest possible BER in an advanced application (**Figures 4** and **5**).

3.3 Architecture of SDR

The Antenna should possess the following characteristics for operation of SDR.

i. It should be flexible to tune to several bands.

ii. Interference rejection

iii. Beam forming Capability (**Figure 6**).

RF Front End has two basic functions:

i. TX/RX signal at various operating frequencies.

ii. Change to/from IF.

The above two process [13] of operation is based on direction of TX/RX mode.
In transmission mode, digital samples are converted by DAC and then mixed with RF frequency and Modulated, then transmitted through Antenna.
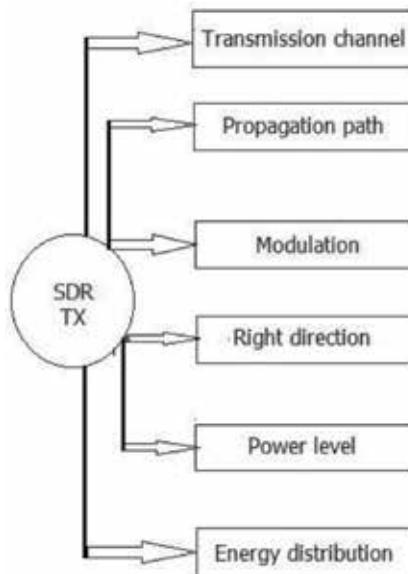
Figure 4.
SDR transmitter.
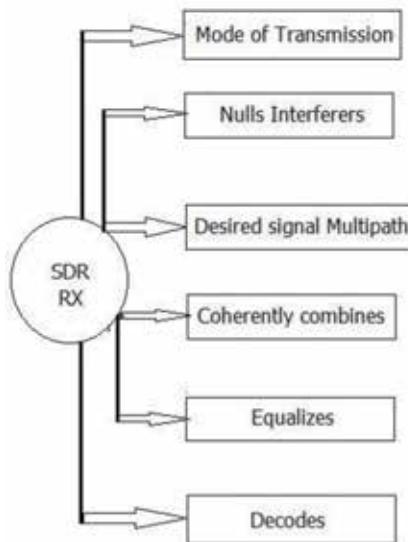
Figure 5.
SDR receiver.
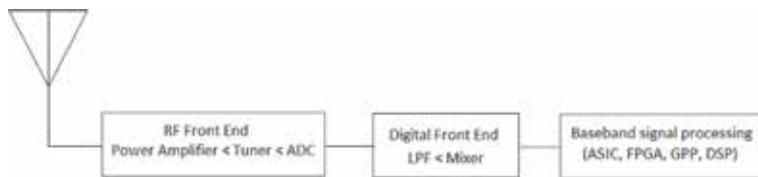
Figure 6.
Block diagram of SDR receiver.
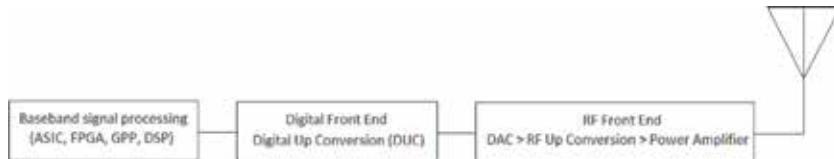
Figure 7.
Block diagram of SDR transmitter.

In receiving mode the antenna is connected to matching/selection circuitry, for reducing the noise it is passes through LNA then to mixer in order to down convert to IF (**Figure 7**).

Digital Front End: To synchronize the TX/RX part to Baseband part it requires conversion of sampling. The DUC in transmitter side translates Baseband signal to IF signal. On Receiver side, DDC which includes Mixer will extracts the baseband signal from ADC and forwards it to signal processing block.

3.4 Introduction to complex signals

In order to analyze and design digital communication systems, it is essential to understand the representation of a signal. A signal is composed of sum of its in-phase (I) and Quadrature (Q) components. The data input to the SDR is complex, which includes I and Q signals.

The Local sampling clock is used as the basic reference in Digital Signal Processing (DSP), therefore it relies heavily on I and Q signals for processing. If real signals (cosine) are shifted from modulated signal to baseband signal, we get sum and difference of frequencies. One key operation in signal processing is the shifting of a signal spectrum from one center-frequency to another. The basis of all the frequency translations lies in multiplying a signal with a complex exponential, generally referred to as complex or I/Q mixing.

3.4.1 Components of sine wave

Sinusoidal signals can be represented as time varying complex numbers Amplitude and Phase (Polar coordinates).
I and Q (rectangular coordinates).
I = In phase (real).
Q = Quadrature (imaginary).
The time domain waveform x (t) of a complex signal is

$$X(t) = xi(t) + jxq(t) \tag{1}$$

$$V(t) = A \times \mathrm{Sin}(2\pi ft + \varphi) \tag{2}$$

Where A is the Peak Voltage, f is the frequency, t is the time, is the phase shift.
The two signals when they are 900 apart in phase are said to Quadrature (1/4 cycles).
Ex: Cosine and Sine waves are in Quadrature.
The Amplitude of the In-phase signal I

$$I \times \cos(2\pi ft) \tag{3}$$

The Amplitude of the 90^0 shifted signal φ

$$\varphi \times \sin(2\pi ft) \tag{4}$$

While Adding the Quadrature Signal, when I and φ vary identically, the Amplitude of the Sum varies (**Figure 8**).
When I and φ vary differently, the phase of the Sum varies. Therefore, $I(t)$ and $\varphi(t)$ variation result in Amplitude, phase and Frequency modulation of the sum. I (t) and $Q(t)$ signals can be generated to result in any modulation such as AM, FM, PM, SSB, BPSK,QPSK, 16QAM, etc. I and Q signals can be easily generated and analyzed in software and processed through ADC and DAC (**Figure 9**).
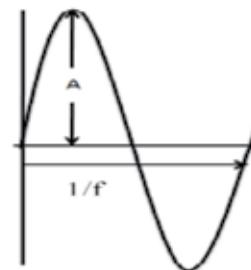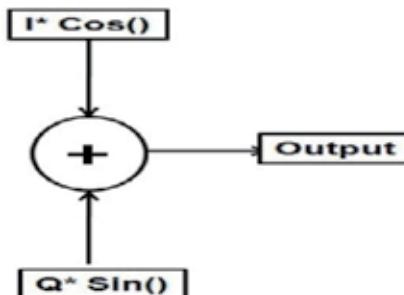
Figure 8.
I and Q signals and a sinusoidal signal.

Nyquist frequency is twice highest frequency of the signal. For example: in analog signal processing, most frequently used frequency is 455 kHz. It requires 910 Ks/s to sample in digital processing, if the signal bandwidth is only 10 kHz. Whereas I and Q, sampling requires only 20 Ks/s.

I and Q allow discriminating of positive and negative frequencies.

$$\text{If:} \qquad H(f) = a + jb \tag{5}$$

$$\text{Then:} \qquad H(-f) = a - jb \tag{6}$$

Representing the basic characteristics of a signal with I and Q:

$$\text{Amplitude } A(t) = \sqrt{I^2(t) + Q^2(t)} \tag{7}$$

$$\text{Phase } \varphi(t) = \tan^{-1}\left[\frac{Q(t)}{I(t)}\right] \tag{8}$$

$$f(t) = \frac{\partial \phi(t)}{\partial t} = \frac{I(t)\frac{\partial Q(t)}{I(t)} - Q(t)\frac{\partial Q(t)}{I(t)}}{I^2(t) + Q^2(t)} \tag{9}$$

The Traditional FM equation

$$x_{FM}(t) = \cos\left(\omega_c t + k \int x_m(t)dt\right) \tag{10}$$

The analytic equation

$$x_{FM}(t) = I(t)\cos(\omega_c t) + Q(t)\sin(\omega_c t) \tag{11}$$

Equations using with I and Q representation for Modulation Techniques.

$$\text{AM}: x(t) = \sqrt{i^2(t) + q^2(t)} \tag{12}$$

$$\text{SSB}: x(t) = i(t) \tag{13}$$

$$\text{FM}: x(t) = \frac{1}{\Delta t} = \tan^{-1}\left[\frac{i(t)q(t-1) + q(t)i(t-1)}{i(t)i(t-1) - q(t)q(t-1)}\right] \tag{14}$$

$$\text{PM}: x(t) = \tan^{-1}\left[\frac{q(t)}{i(t)}\right] \tag{15}$$
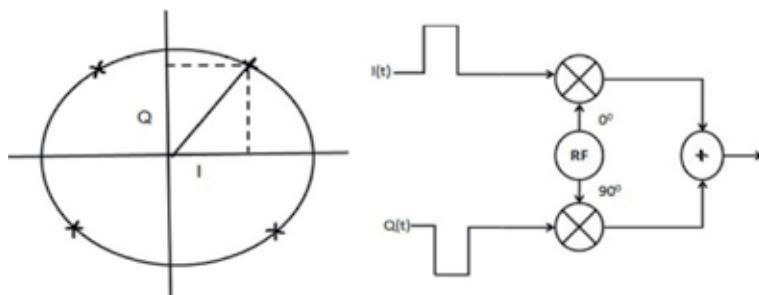
Figure 9.
Phasor diagram and quadrature signal generator.
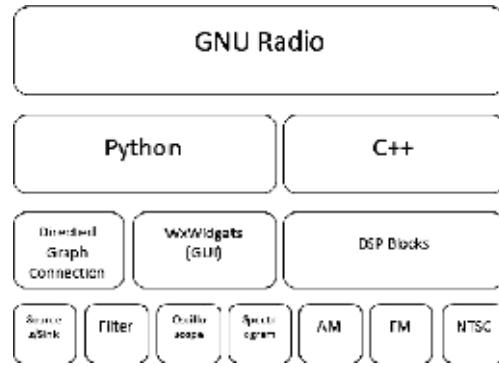
Figure 10.
Basic structure of GNU radio.
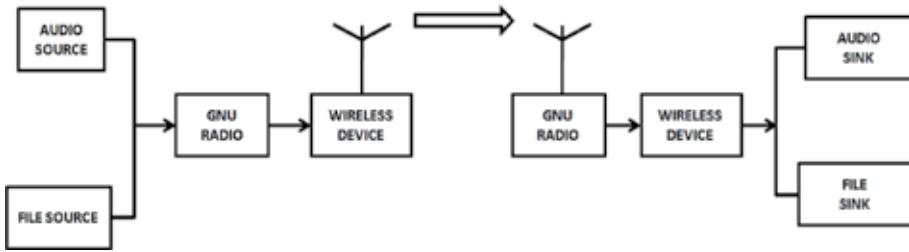
Figure 11.
GRC architecture for transmitter and receiver.

3.5 GNU radio

C++ and Python are the two major layers that are used in GNU Radio. For small signal processing task code are written in C++. The high level codes are written in Python used as an interpretable language, it mainly perform the work of connecting the various signal blocks to make a signal Flow graphs. Python does not require any additional compilation time so it is mainly used for the rapid functioning of the flow graph (**Figure 10**).

The GNU Radio software provides a set of signal processing blocks that can be aggregated to build flow-graphs shown in **Figures 10** and **11**. These blocks are written in C++ and run in Python, which brings several advantages, such as easy instantiation and connection of existing blocks and easy GUI (Graphical User Interface) creation, as shown in **Figure 10**, Around 250 blocks in GNU Radio include various applications like simple mathematical operation blocks, modulators/demodulator blocks, channel coding blocks, voice codec and others. Input/output blocks are the Special classes of blocks.

In this system, Host Computer is any normal Laptop/CPU with GNU Radio installed over it or running with the help of live USB environment is shown in **Figure 11**.

4. Signal processing blocks in GNU radio

It Consist of complex flow graph that includes modules and low level algorithm. The basic signal processing functions such as filters and channel coding are structured in C++ [14] called to be low level algorithm or module. With the interface of wrapper SWIG (i.e., Simplified Wrapper and Interface Generator) they are generated as python modules.

With the help of python, the generated blocks are used to construct a flow graph model shown in **Figure 12**. The application is built on python program that provides python framework. The python frame work creates a simple scheduler to run blocks in a sequential order for every iteration and it is responsible for communication of data through module buffers.

Signal Source: The head (start) of the flow-graph is the source. For instance, Osmocom Source, USRP source, and file source are common types of source blocks.

Signal Sink: Sink is used at the end of the flow-graph. For instance, USRP sink, file sink are common types of sink blocks.

Flow-graph: The application is based on a flow-graph. Each flow-graph consists of intermediate processing blocks along with single source and sink blocks. We can have multiple flow graphs within a single application.

Scheduler: It is created for each active flow-graph, which is based on steady stream of data flow [15] between the blocks. It is responsible for transferring data through the flow-graph. It monitors at the I/p and O/p buffers of each block for sufficient data so as to trigger processing function for those blocks.

Three important stages are observed in a basic Radio. They are Front End, IF stage and Demodulator stage. Similarly GRC flow graph consist of following blocks like Source, Sink Filter, and Demodulator/Modulator.

Front end stage or Source Block is required for radio, whose function is to tune frequency, sample rate, applying the RF gain etc. These three are essential after receiving the signal from Antenna. In GNU Radio this front end stage is provided by USRP Source block, Osmocom Source block and RTL-SDR source block.

Intermediate stage will fulfill the function called sample rate conversion. Which consist of steps like up-sampling (interpolation), down-sampling (decimation), Low pass filter, Band pass filter is used.

In **Demodulator stage** Appropriate Demodulator [16] (AM Demodulator, NBFM Receiver, and WBFM Receiver) is used to recover the signal. Sometimes amplifier is used to support the end device.

Graphical User Interface (GUI) in GNU Radio displays the signals, in Time/ Frequency, scope to observe the waterfall, histogram scope, constellation scope for observing the digital modulation schemes. The following blocks called Chooser block, variable, notebook, slider to provide tuning, selecting the parameters.

Sampling rate conversion (SRC) [17, 18]: SRC must synchronize the Clock rate of ADC and baseband signal processor. Main reason of processing of sample rate conversion in reconfigurable radio by SDR software is to place the ADC as close as possible to antenna without changing underlying hardware, to emulate any signal of different frequencies.

Reconstruction after re-sampling process: let a signal at f1 is to be converted to f2 rate the popular method is to perform DAC and resample the reconstructed signal (**Figure 13**).

During resampling process spectral repetition is caused, in turn spectral overlap occurs which is called Aliasing. This can be avoided by reconstruction filter $h_c(t)$ to band limit the signal. To suppress aliasing and imaging effect the filters along Decimators and Interpolators are used respectively (**Figure 14**).

Decimation/Sample rate reduction: Anti-aliasing filter followed by down-sampling is known as decimation.

Figure 12.
Basic structure of GNU radio flow graph.
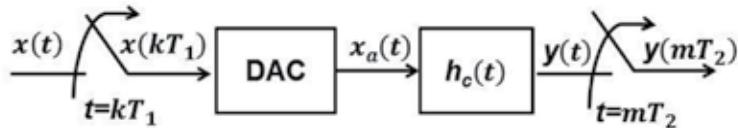
Figure 13.
Sampling rate conversion.
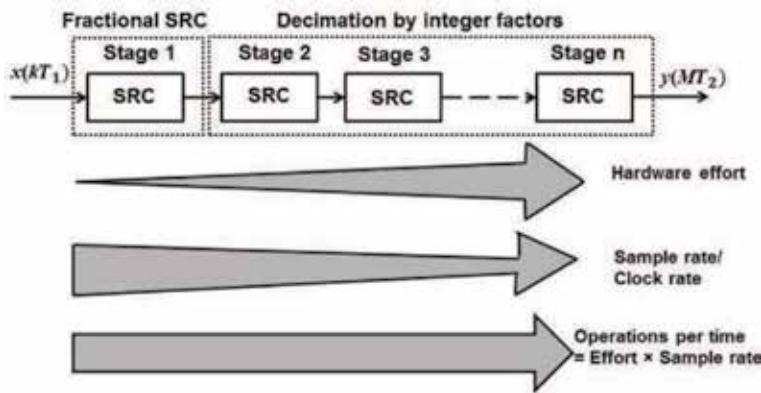
Figure 14.
Multistage SRC.

Interpolation/Sample rate increase: Up-sampler followed by anti-imaging filter (Interpolating Filter) is called Interpolation. The hardware effort increases as the sampling rate decreases from first stage to last stage in a multistage SRC system.

Rational factor SRC: when L/M both are positive integers to provide combined effect of reducing anti imaging and anti-aliasing it is called Rational Factor.

The benefit of using Multistage SRC is computational complexity is reduced and simplified filter design.

4.1 Analog modulation techniques

4.1.1 Amplitude modulation (AM)

A message or signal can modify a radio wave by changing its amplitude [19], phase or frequency. Amplitude modulation changes the amplitude of a radio wave in concert with changes in the message. Some real time examples such as Sound signal at the speaker end, or intensity of light for television pixels, etc. Implementation of Amplitude Modulation is shown in **Figure 15**.

The data flow graph has been designed by considering the following equations. The carrier wave of frequency f_c and amplitude A is given by:

$$c(t) = A \times \mathrm{Sin}\left(2\pi f_c t\right) \tag{16}$$

where, f_c = 10 kHz and A = 1.
The modulation waveform of frequency f_m and amplitude M is given by:

$$m(t) = M \times \mathrm{Sin}\left(2\pi f_m t\right) \tag{17}$$

where, f_m = 1 kHz and M = 0.7.

The amplitude-modulated wave is obtained by considering the following equation:

$$y(t) = [1 + m(t)] \times c(t) \qquad (18)$$

4.1.2 Generation of AM using GNU radio

The following are the Requirements for implementation of AM:

The two Signal Sources are required one for Message and other for Carrier signal, and pass the signals through multiplier (product) to generate AM modulated signal (**Figure 16**).

At the receiver side there is need for Sample Rate Conversion due to the following reason:

Sampling-rate conversion can be accomplished by L-fold expansion, followed by low-pass filtering and then M-fold decimation. It is important to emphasis that the interpolation should be performed first and decimation second, to preserve the desired spectral characteristics of x[n]. Furthermore by cascading the two in this manner, both of the filters can be combined into one single low-pass filter. Demodulator is followed by the filter to recover the signal.

In GNU Radio all the modulation techniques are implemented in Software, and also by applying mathematical operations to streaming data for low pass digital filters, digitally generated Oscillators frequency Synthesis, signal processing algorithms are also implemented as software.
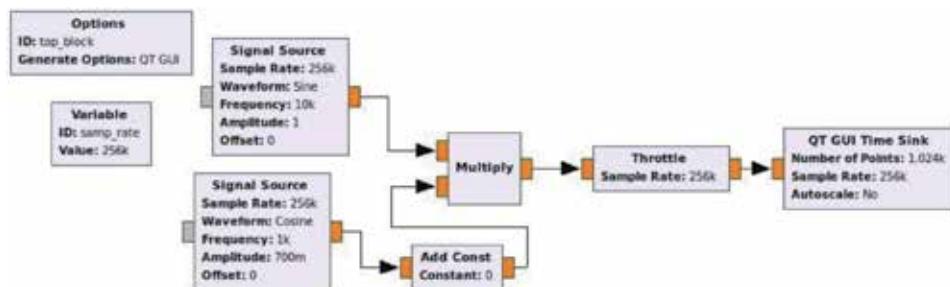
Figure 15.
GRC flow graph for generating amplitude modulation signal.
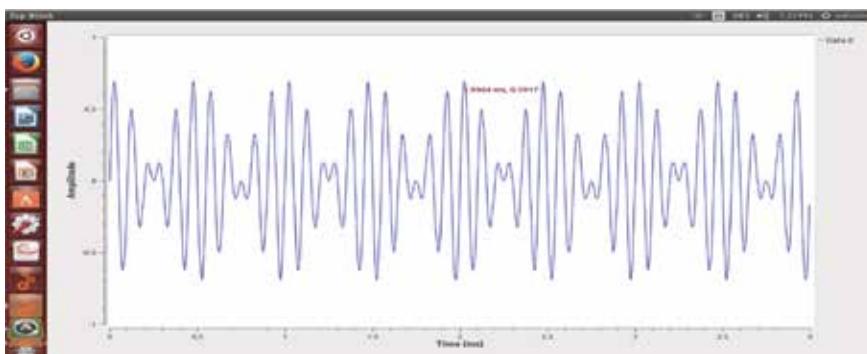
Figure 16.
Waveform of AM modulator.

4.1.3 Functions of the blocks in flow graph

Signal Source: This block facilitates to define the sample rate, output waveform shape (pulse, square, triangle).If the frequency in the source is set higher than half of the sample rate, aliasing will occur. Offset option is also provided to add to generated waveform.

Multiply: This Block Multiplies input stream by a constant.
Output = Input * Constant Scalar/Vector (element-wise if vector).

Add Const: To obtain modulation, change the amplitude levels of modulated signal (e.g. from 0 and 1, to 1 and 2). It can be done by adding constant value to modulated signal (Component "add const").

Throttle: Limits the data throughput to the specified sampling rate. This prevents GNU Radio from consuming all CPU resources when the flow graph is not being regulated by external hardware (i.e. audio source/sink or USRP source/sink). Specifies the data type of the input and output, the sample rate to limit the flow graph and the vector length for vector processing.

Complex Mag ^ 2: These blocks convert a complex signal to the magnitude of the complex signal and the phase angle of the complex signal. The argument v_{len} is the vector length; always use the default value 1.

Rational Re sampler: This block is Combined Interpolator and Decimator blocks. This block is used to convert from one sample rate to another as long as they can be related by a ratio: *Fs_out = Fs_in* * Interpolation/Decimation. All following block in the flow graph can able to expect the output sample rate.

Audio Sink: After completing all the processing, to play the signal through the speaker, use the audio sink. Audio sink will output the data to the sound card at the sampling rate of sampling rate.

WX Slider: Sliders are used to adjust the frequency. Sliders are used in flowchart for adjusting carrier frequency(10–100KHz), for adjusting modulating frequency (10-1000 Hz), Volume control, Gain Control, adjusting Amplitude of signals and also used to adjust the filter cutoff frequency.

Simple AM transmitter and receiver are shown in the flow graph **Figure 17**. To transmit an audio signal of 48Khz, before the signal need to be up converted by choosing appropriate L and M values for the required sample rate conversion. However noise is predominant in this type of modulator (**Figure 18**).

SSB Modulation: Single-sideband modulation (SSB) is an improvement over amplitude modulation which uses transmitter power and bandwidth more efficiently. The data flow graph of the SSB modulator are designed using the following equations

$$s_{LSB(t)} = s(t) \times \cos\left(2\pi f_c t\right) + \hat{s}(t) \times \sin\left(2\pi f_c t\right), \text{for lower side band} \qquad (19)$$

$$s_{USB(t)} = s(t) \times \cos\left(2\pi f_c t\right) - \hat{s}(t) \times \sin\left(2\pi f_c t\right), \text{for upper side band} \qquad (20)$$

Where s(t) is the information signal, $\hat{s}(t)$ is the Hilbert transform of the data signal and f_C is the carrier frequency

$$s(t) = \cos\left(2\pi f_c t\right) \qquad (21)$$

For upper side band modulator the data flow graph is as shown in **Figure 19** and the output wave form in frequency domain is shown in the **Figure 20**.

4.1.4 GRC flow graph of generation of SSB modulation

Refer the following blocks such as Signal source, Throttle, Multiply and FFT Sink in the AM Modulator.
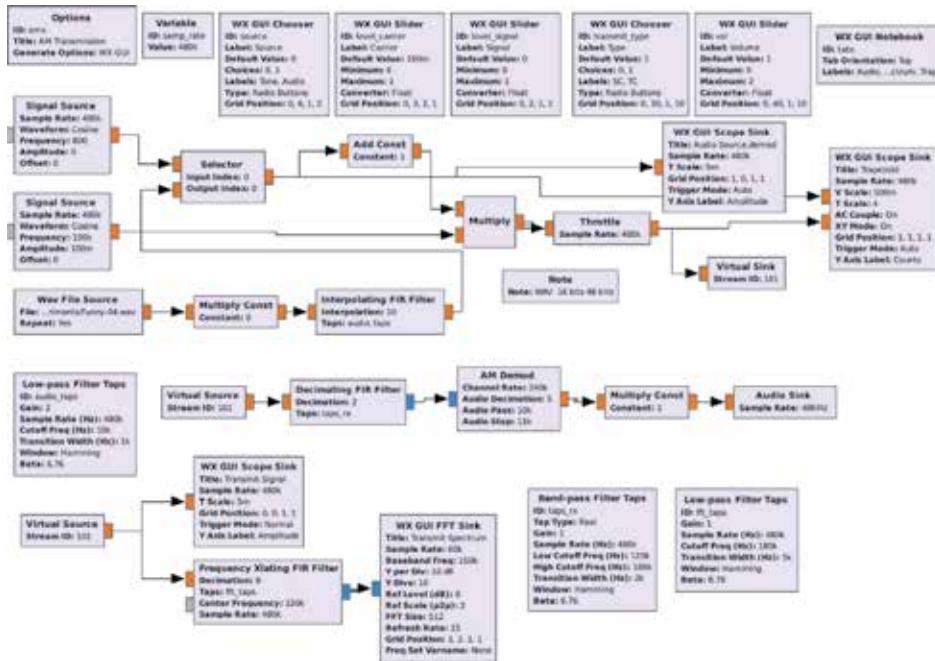
Figure 17.
GRC flow graph for AM transmitter and receiver.
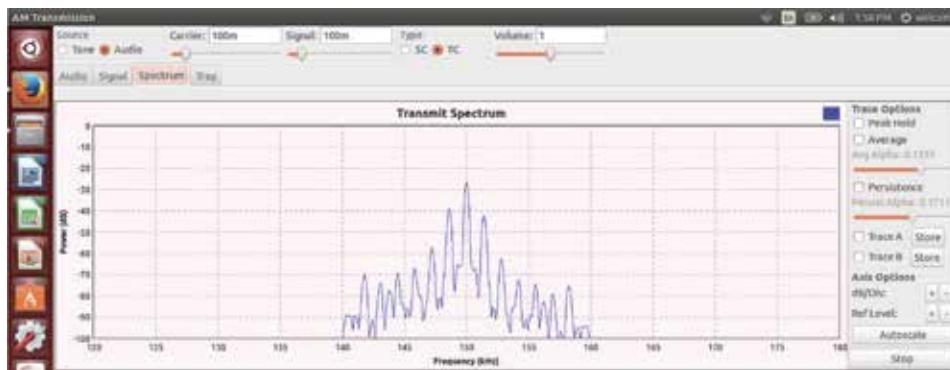
Figure 18.
AM transmitter spectrum.

Hilbert: This block is basically a non-causal filter. For the applied real signal it performs the Hilbert transform.

Complex to Float: This block converts the complex to float, it has 2 outputs, one is real and other is imaginary output.

4.1.5 GRC flow graph for SSB transmitter and receiver

Single sideband modulation is the most suitable choice to be used for analogue voice transmission because of less power consumption and spectrum efficiency. Single sideband modulation is the most suitable choice to be used for analogue voice transmission because of less power consumption as it avoids double bandwidth and spectrum efficiency. Out of three methods to implement single-side band
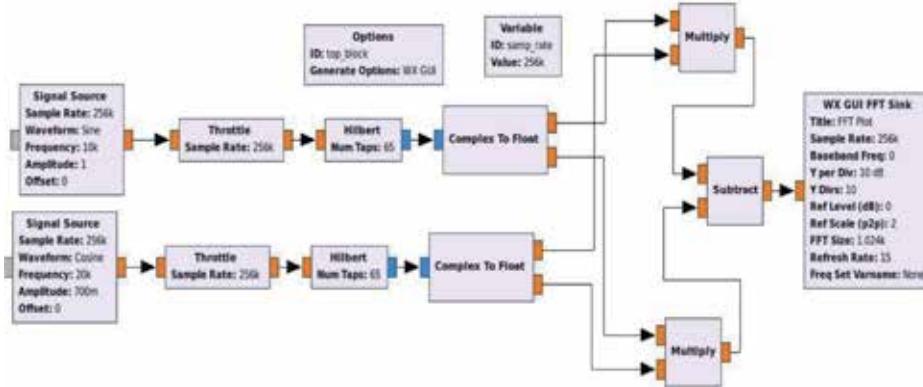
Figure 19.
GRC flow graph signal generation of SSB modulator.
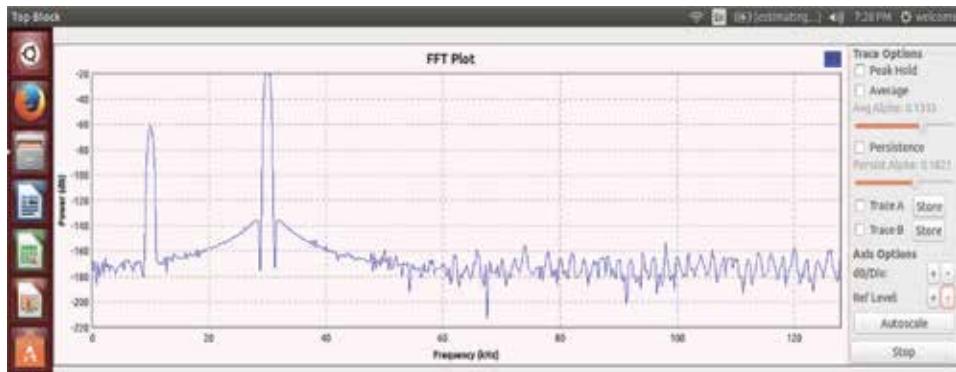
Figure 20.
GRC flow graph for SSB modulation.

modulator, Band-pass filtering method has a simple structure to implement in a digital way and is shown in **Figure 21**.

Frequency Modulation: The instantaneous carrier frequency is varied according to baseband signal is called frequency modulation (**Figure 22**).

Carrier wave $x_c(t) = A_c \cos\left(2\pi f_c t\right)$.
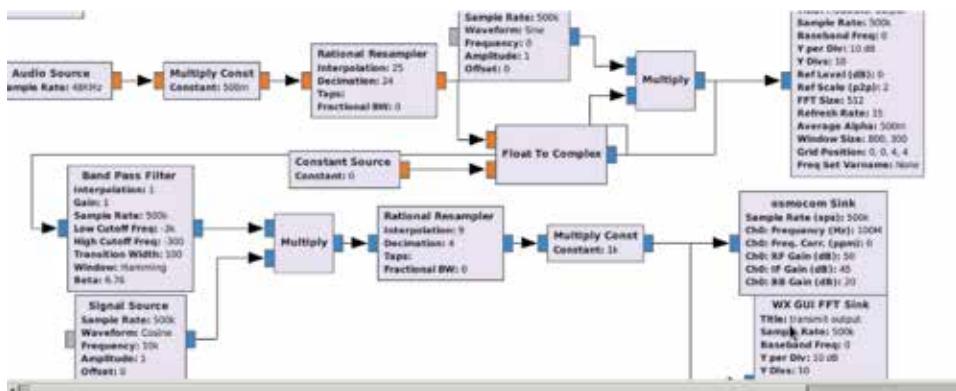
Figure 21.
GRC flow graph for SSB transmitter and receiver.

The modulator output

$$Y(t) = A cos((2\pi f_\Delta \int_0^t x_m(\tau)d\tau$$

Modulation Index $f_\Delta = k_f A_m$.

Where k_f is the sensitivity of frequency Modulation and A_m is the amplitude of baseband signal.

4.1.6 Design of FM modulation and demodulation

VCO plays major role in FM signal generation shown in **Figure 23**. The modulation index determines the volume and sound frequency of the signal. One of the major advantage of FM is the signal can be recovered from sideband (Bessel function), FM demodulator is used to recover the baseband signal. For entertainment broadcasting applications WBFM is used, whereas NBFM is used in communication like police wireless, ambulance etc. (**Figure 24**).

4.2 Digital modulation techniques

4.2.1 ASK modulation

It is the digital to analog conversion technique. Amplitude of carrier varied according to amplitude of message signal. For binary n-1 it is called as Binary ASK (BASK) Or On off Keying Technique (OOK) [20]. Bandwidth is directly proportional to baud rate of Message signal m (t).
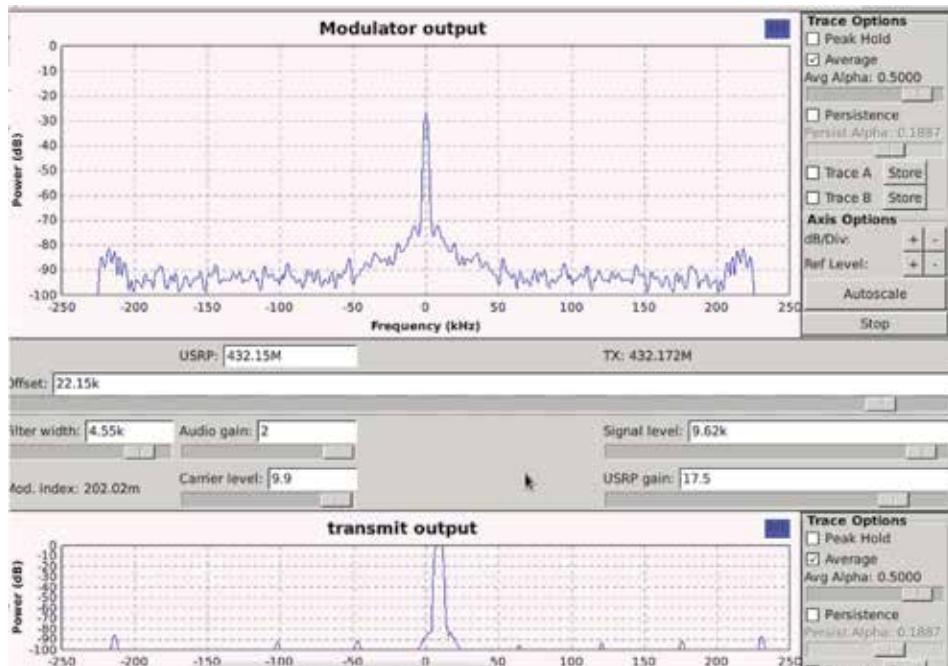
Figure 22.
Waveform of transmission of SSB signal.
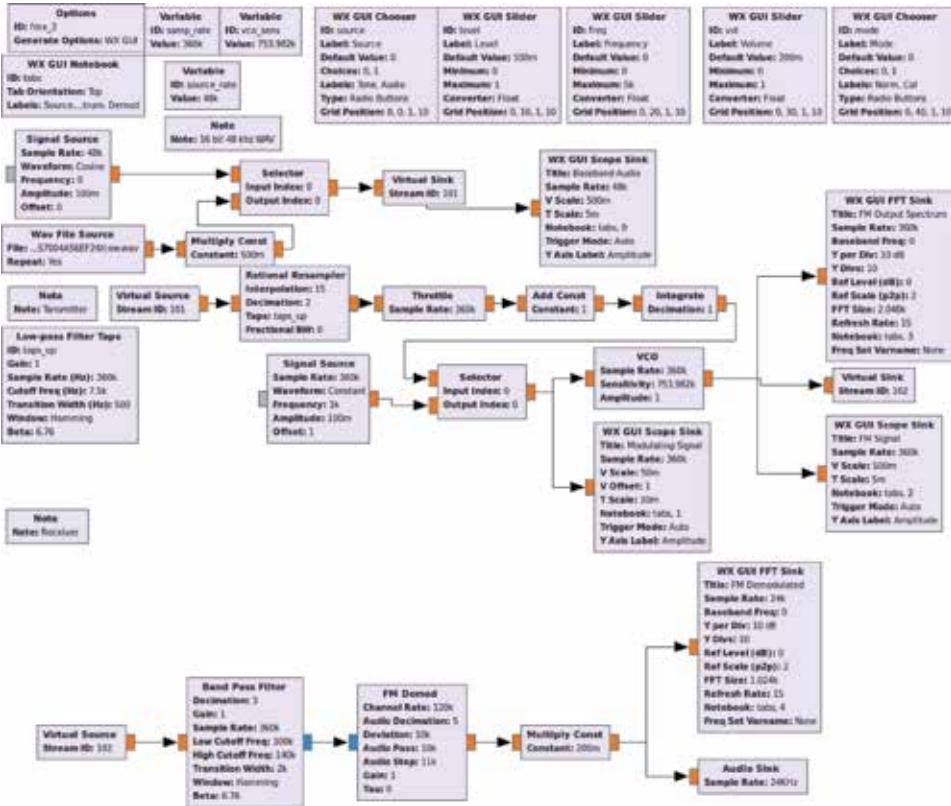
Figure 23.
GRC flow graph of FM modulator and demodulator.

$$B \propto r \tag{22}$$

$$B = (1+d)r \tag{23}$$

Where d is the factor for modulation and filtering process (Range 0 for ideal case then B = r, and 1 for worst case then B = 2r) (**Figure 25**).

Where baud rate = data rate /No of bits required for sample r = R /n

$$B = (1+d)R/n \tag{24}$$

4.2.2 ASK demodulator

See **Figure 26**.

4.2.3 GRC flow graph of ASK modulation and demodulation

ASK/OOK modulation: Blocks such as Signal Source, Multiply, Throttle, Add Const blocks been discussed in AM modulation flowchart.

Float to Complex: These blocks are like 'adapters', used to provide the interface between two blocks with different types. If there are two inputs, they serve as the real and imaginary parts of the output signal respectively. To change the data type from floating point (real) to complex this block is deployed (**Figures 27** and **28**).

Channel Model: The Transmission channel model is used to best reproduce the behavior of the signal during transmission, in the form of electromagnetic wave. It is a model of the channel in which the noise is constant and spectral power density
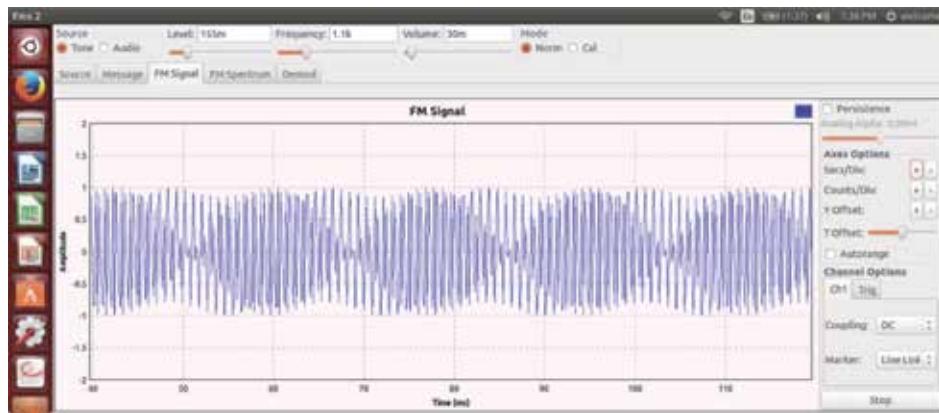
Figure 24.
Waveform of FM modulation.
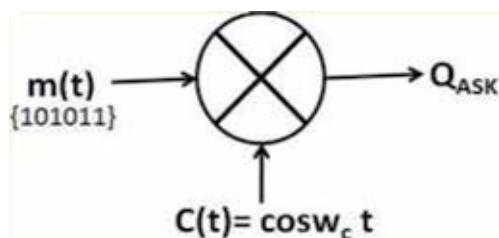
Figure 25.
ASK modulator block diagram.
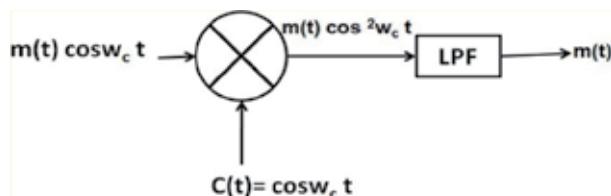
Figure 26.
ASK demodulator block diagram.

(watts/Hz) and a Gaussian distribution of amplitude. This model is represented by the component "Channel Model".

Frequency Xlating FIR Filter: These blocks are used as FIR Filters combined with frequency translation. Reason is that in the FPGA there are digital down converters (DDC) translating the signal from IF to baseband. To select the narrow bandwidth channel, it is necessary to perform decimating FIR Filter in order to perform frequency down sampling. From the receiver in order to performs frequency translation on the signal Frequency Xlating FIR filter is used.

Clock Recovery MM: Mueller and Müller(M&M) clock recovery is the name of one of the signal processing blocks in GNU Radio. Its task is to recover samples from a signal with the same frequency and phase as those used by the transmitter. On modern SDR systems the software layer does not have the ability to control directly the sampling rate on the ADC, usually residing in a different piece of hardware. M&M was designed in 1976 to be implemented in hardware and connected to the ADC clock reference.
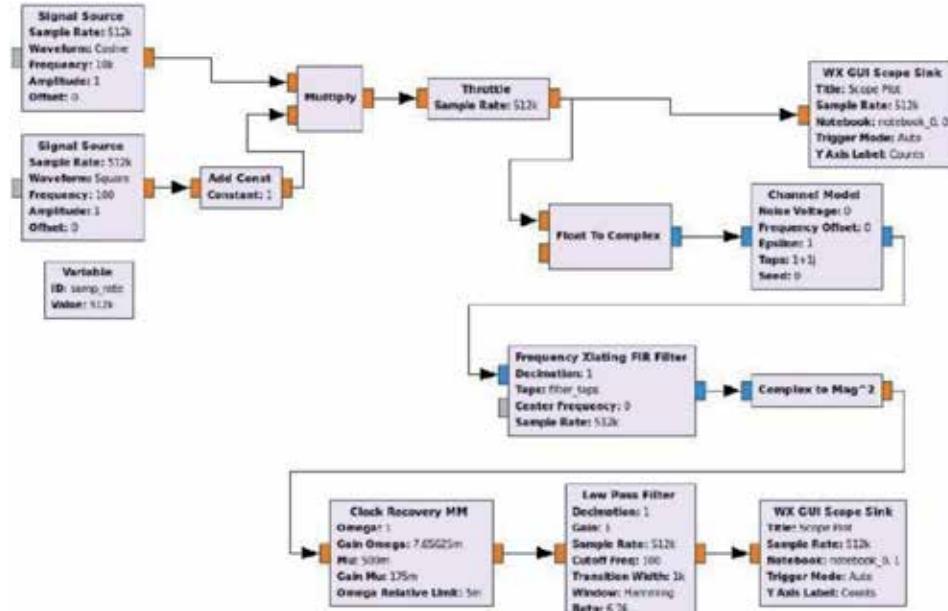
Figure 27.
GRC flow graph for ASK modulation and demodulation.
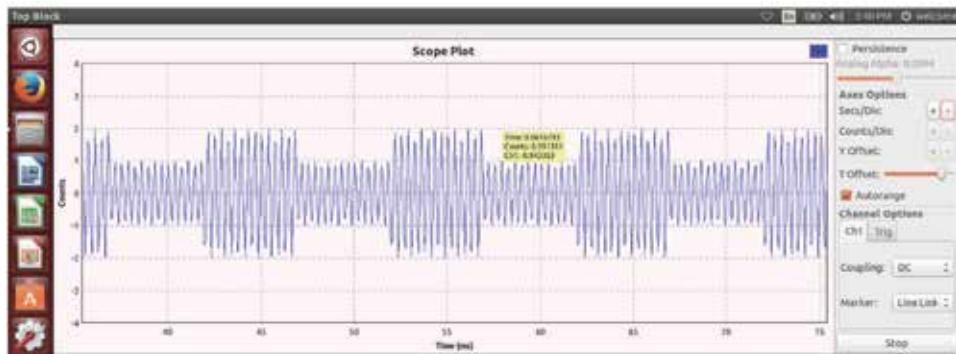
Figure 28.
Waveform of ASK modulation.

Low pass Filter: Filters are used to separate the mixed signals, extract a specific band of interest, avoid undesired effects (aliasing, imaging, etc.), and restore distorted signals. The unique features LPF block representing a digital filter in GNU Radio is that it is implemented in software, variable performances depending on the processing resources, change their properties relatively easily, can adapt to any application.

WX_Scope_Sink: This block displays the results of signal modulation on the scope.

ASK allows for two amplitude levels, two symbols 0 and 1.the transmission rate is determined by Symbol rate and Bit rate. The rate at which amplitude was changed per second is measured in Bauds. ASK is the simple modulation technique which offers high bandwidth efficiency used at radio frequencies to transmit more codes. It offers low power efficiency and very susceptible to noise interference (**Figures 29–32**).

4.2.4 FSK modulation

Frequency of carrier signal is varied w.r.t amplitude of message signal [21].
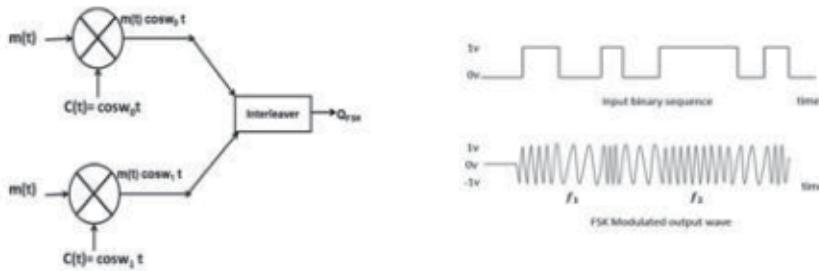
Figure 29.
FSK generator block diagram and waveforms.

$$\varphi_{FSK} = m_0(t) \cos w_0 t \qquad (25)$$

$$\varphi_{FsK1} = m_1(t) \cos w_{c1} t \qquad (26)$$

$$m(t) = \left\{ \begin{array}{l} 1 - f_{c1} \\ 0 - f_{c0} \end{array} \right\} \qquad (27)$$

$$BW_{min} = r + r = 2r \qquad (28)$$

Total Band width of FSK is $(1 + d)r = 2\Delta f$ for minimum Bandwidth = 0, substituting we get $2\Delta f = r$(r is the baud)

4.2.5 FSK demodulator

See **Figure 30**.

4.2.6 GRC flow graph of generation of FSK modulation

See **Figures 31** and **32**.

4.2.7 FSK modulation: blocks and its functions in detail

The above flow graph shown in **Figure 20** and **Figures 27** and **28** includes the FSK transmitter and the receiver. The output frequency is 433 MHz in ISM band which is a free space to transmit. **Figures 27** and **28** shows the GRC Flow Graph of the transmitter that is described as follow.

In the above block diagram **variable Blocks** are used with sample rate 512 k Samples per second.
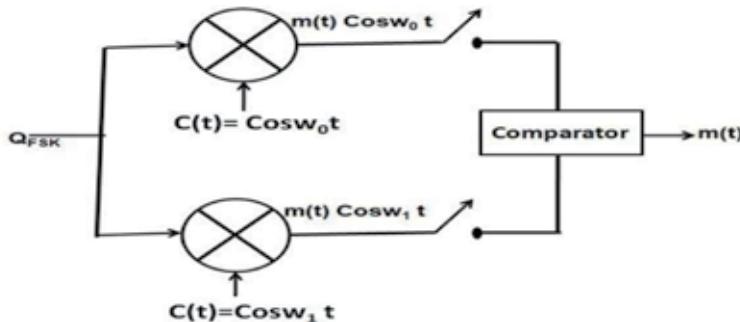
Figure 30.
FSK demodulator block diagram.
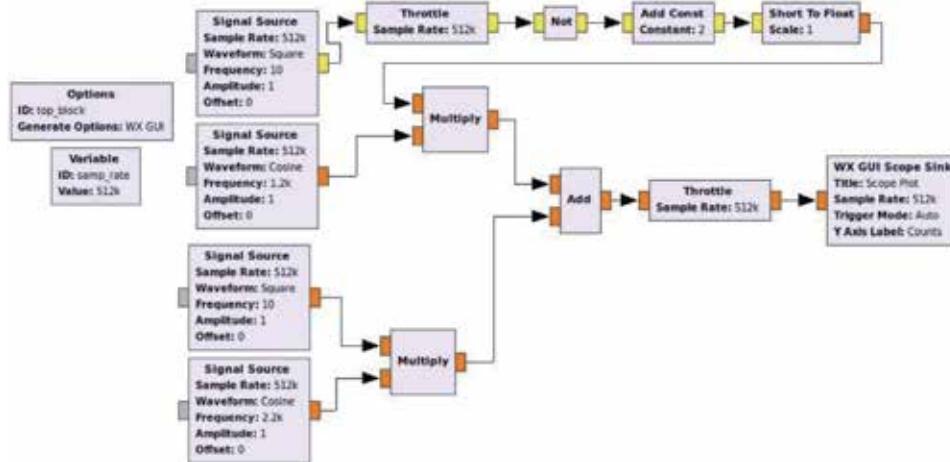
Figure 31.
GRC flow graph for FSK modulator.
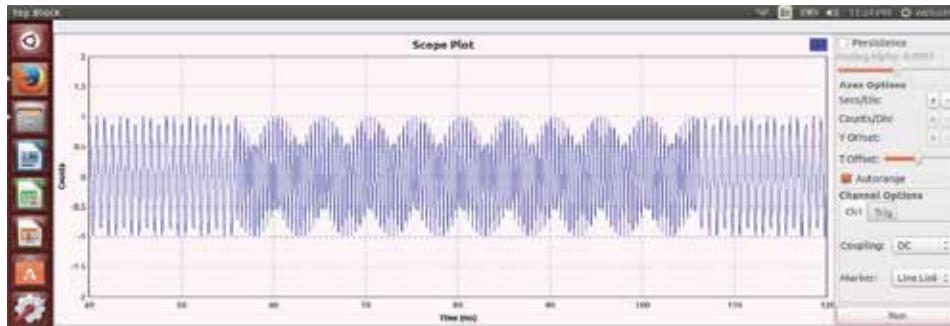
Figure 32.
Waveform of FSK modulator.

Throttle block with the sample rate proposed to keep the sample rate constant when the hardware attached to a computer.

Osmocom sink used to link the interface with HackRF one to control the output with the sample rate of 1 M, 440 M frequency, RF gain of 10 dB, IF gain of 20 dB and BB gain of 20 dB.

Add Block Implements the function out = in + Constant.

This block Specifies the constant to add.

4.2.8 GRC flow graph of GFSK modulation and demodulation

See **Figures 33** and **34**.

FSK Transmitter and Receiver operated in ISM band, The digital File source with the samples are first up sampled by using Sampler and to keep the sample rate constant when the hardware is not attached to a computer throttle block is used. The access code and preamble can leave blank, in Packet encoder but if new access code is created set the packet decoder codes so that match perfectly for the decoder to pass the data for synchronization. GFSKMod was for modulating. To convert complex signal Quadrature Demod Gain was set 1.

FSK has higher immunity to noise due to constant envelope, but the BER (Bit Error Rate) performance in AWGN channel is worse compare to PSK modulation.
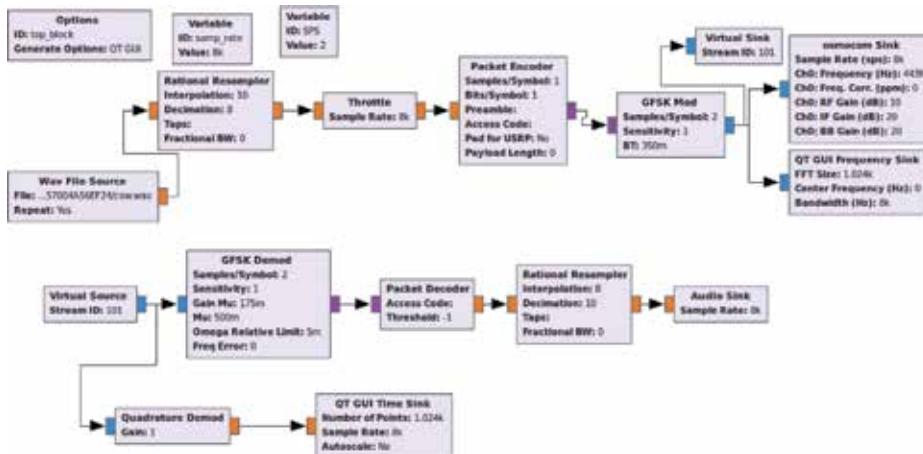
Figure 33.
GRC flow graph of GFSK modulation and demodulation.
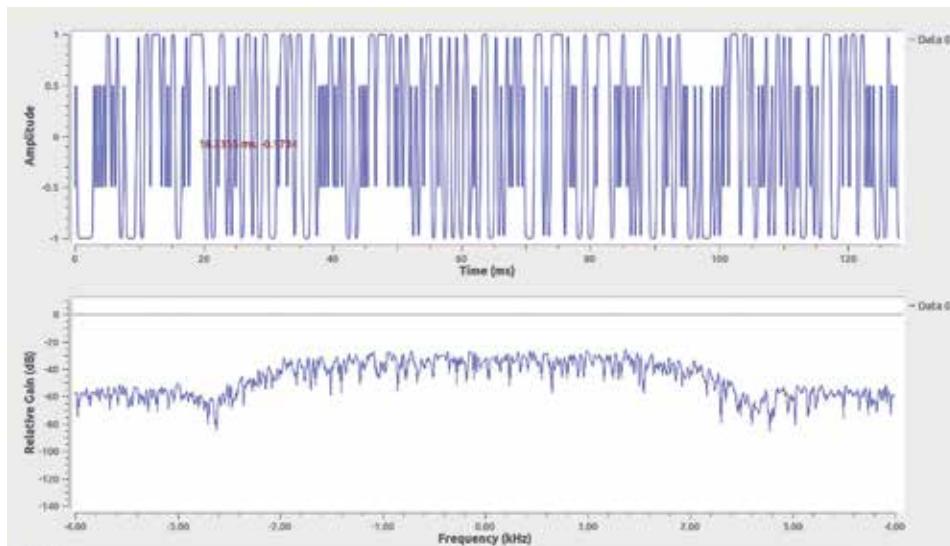
Figure 34.
Waveform of GFSK modulation and demodulation.

4.2.9 PSK modulation

Carrier phase is varied w.r.t amplitude of message signal [22].

Binary 1 = Phase Q = 0°

and 0°= Phase Q = 180°

For BPSK total levels is L = 2, n = 1 bit i.e. Phases Q = 0° and 180°.

For Multilevel PSK let L = 4, n = 2 bit 360/4 i.e. Phases Q = 0°, 90°, 180°, and 270°.

For Multilevel PSK let L = 8, n = 3 bit 360/8 i.e. Phases Q = 0°, 45°, 90° 135°, 180°, 225°, and 270°.

PSK is better than ASK and FSK Modulation. Its bandwidth is better than FSK, Less Immune to Noise, and Data rate is better than FSK Non coherent detection is not possible and costly (**Figure 35**).
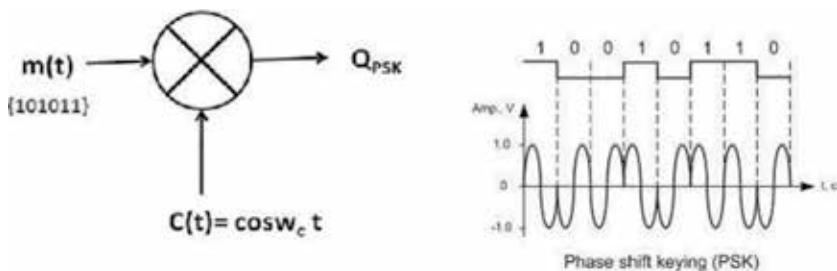
Figure 35.
PSK modulator and waveforms.

$$Y(t) = \varphi(t)_{PSK} \times \cos W_C t$$
$$= \pm m(t) \cos^2 W_C t$$
$$= \frac{m(t)}{2} \pm \frac{m(t)}{2\cos^2 W_C t}$$

4.2.10 PSK demodulator

See **Figure 36**.

4.2.11 GRC flow graph of generation of BPSK modulation

See **Figures 37** and **38**.
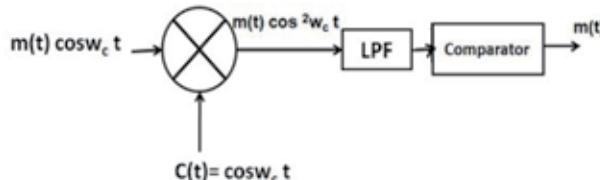
Figure 36.
PSK demodulator.
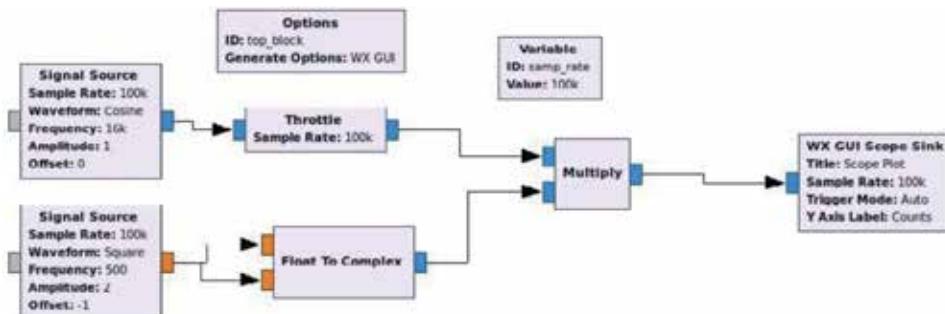
Figure 37.
GRC flow graph for generation of PSK modulation.
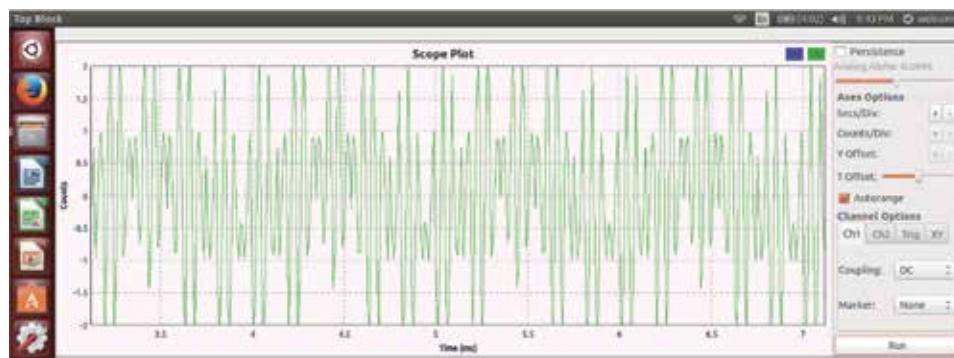
Figure 38.
Waveform of PSK modulation.

4.2.12 PSK modulation: blocks and its function in detail

In this modulation scheme, the data bits are mapped to the phase of carrier frequency and these two are separated from each other by 180 degrees. Map 0's and 1's to φ and 0 and then add them to the phase of carrier frequency. One can view adding 0 and φ to the phase of the carrier wave, same as changing its direction without changing its magnitude.

Random source generates 0's and 1's which are converted to float. Then to map these 0's and 1's to −1's and 1's respectively multiply it with 2 then add −1.([0 1] * 2) −1 gives us −1 or +1. Now feed the output in QT GUI sink block, where you'll see two dots at −1 and +1 on IQ-plot, same as one would expect from BPSK transmitter.

QPSK Modulation: It is an extension of the phase shift keying (PSK). As the phase of a signal is in quadrature or 90 degrees to another one it is called as quadrature. This reduces the bit rate and hence reduces the bandwidth of the channel.
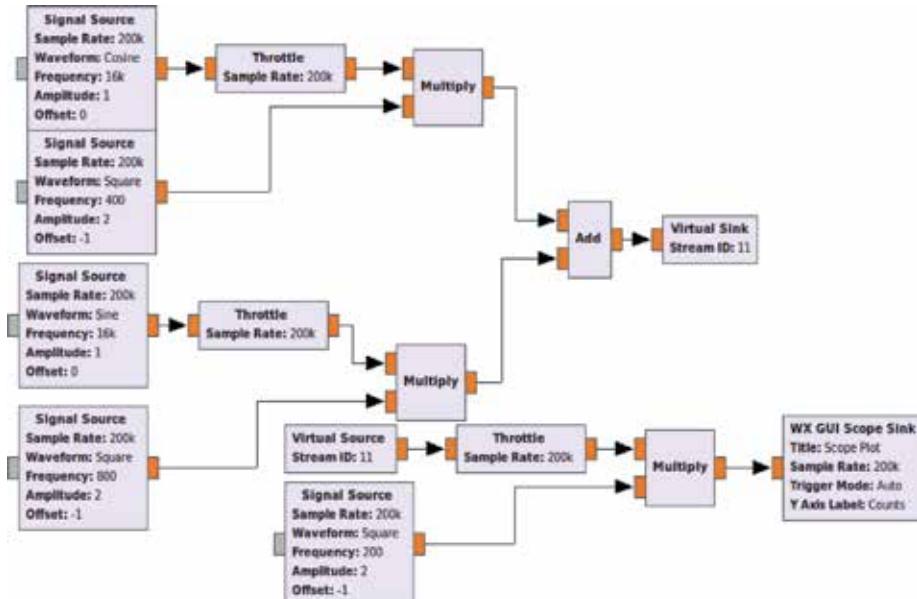
Figure 39.
GRC flow graph of QPSK signal generation.
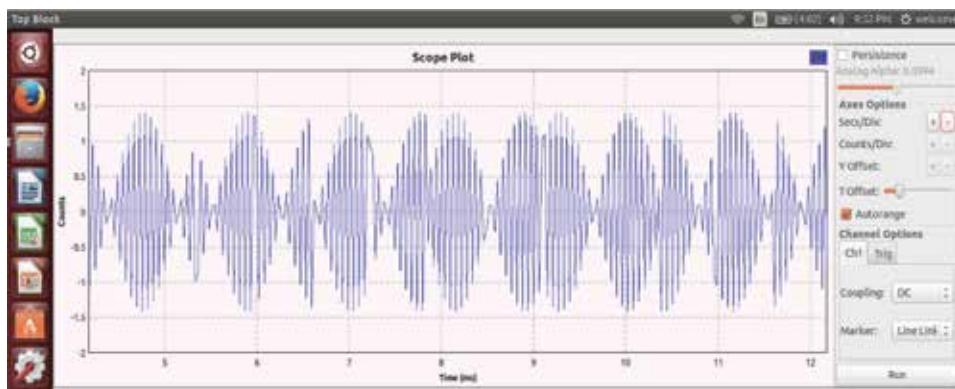
Figure 40.
Waveform of QPSK modulation.

4.2.13 GRC flowgraph of QPSK signal generation

See **Figures 39** and **40**.

4.2.14 Adaptive modulation in cognitive radio

The specialty of approaching cognitive radio is Adaptive Modulation [23], which has the ability to select the required modulation scheme necessary in the Communication system. The steps followed are. For performing this a)track the network behavior d)decision on the change of modulation techniques (ASK,BPSK, QPSK, and DPSK)during on-line transmission, while maintaining the real time objective. To scan the channel state to avail opportunistic spectrum access (OSA) and spectrum sharing.

Initial test is conducted to evaluate the performance of various modulation techniques to develop training sequence in multi-user cognitive fading environment. The performance improvement and the Spectral Efficiency both are achieved using Adaptive Modulation in cognitive radio.

5. Conclusions

In advanced communication system there is a need for smart radio which can aware of environment and select the channel with good signal strength. As a basic course this chapter covered the architecture, characteristic of SDR, and implementation of various modulation concepts using SDR. By experimenting Modulation techniques using GNU Radio, research can be extended by understanding certain real time parameters like channel characteristics and its effect, Fading caused by Multipath, Time delay Spread, Filtering and Tuning, MultiMate Signal processing blocks and its application in flow graph such as Interpolation and Decimation concepts helps to find the filter performance and tuning. It provides Visual Scope to observe Frequency/ Time domain and Constellation plots for observing real time parameters such as SNR, ISI, and BER etc. By applying concepts, with different flow graph and algorithm the SDR explore the communication system in real world scenario.

Author details

Madhuri Gummineni* and Trinatha Rao Polipalli
GITAM School of Technology, GITAM Deemed to be University, Hyderabad
Campus, Hyderabad, Telangana, India

*Address all correspondence to: madhuri.vijay2003@gmail.com

IntechOpen

References

[1] Available from: http://www. sdrforum.org/pages/documentLibrary/ documents/SDRF-06-R-0011-V1_ 0_0.pdf

[2] Khozeimeh F, Haykin S. Dynamic spectrum management for cognitive radio: An overview. Wireless Communications and Mobile Computing. 2009;**9**:1447-1459. Published online 6 January 2009 in Wiley InterScience (www.interscience. wiley.com). DOI: 10.1002/wcm.732

[3] Farrell R, Sanchez M, Corley G. Research article software-defined radio demonstrators: An example and future trends. International Journal of Digital Multimedia Broadcasting. 2009:1-12. DOI: 10.1155/2009/547650. Article ID: 547650

[4] Thomas RR, Zayen B, Knopp R, Maharaj BT Multiband Time-of-Arrival Positioning Technique for Cognitive Radio Systems. 2011 IEEE 22nd International Symposium on Personal, Indoor and Mobile Radio Communications ©2011 IEEE

[5] Alsharoa A, Neihart NM, Kim SW, Ahmed E. Kama Multi-band RF Energy and Spectrum Harvesting in Cognitive Radio Networks. arXiv.org > cs > arXiv: 1804.03533v1 10 April 2018

[6] Zhang X, Guo L, Song T, Xu W, Li Y, Lin J. Dynamic channel allocation supporting multi-service over cognitive radio networks. In: 16th International Symposium on Wireless Personal Multimedia Communications (WPMC), June 2013.

[7] Elhachmi J, Guennoun Z. Cognitive radio spectrum allocation using genetic algorithm. EURASIP Journal on Wireless Communications and Networking. 2016;**2016**:133. DOI: 10.1186/s13638-016-0620-6

[8] Hernandez C, Salgado C, López H, Rodriguez-Colina E. Multivariable algorithm for dynamic channel selection in cognitive radio networks. EURASIP Journal on Wireless Communications and Networking. 2015;**2015**:216

[9] Bharatula S. Efficient channel allocation schemes for cognitive radio networks. International Journal of Scientific and Engineering Research. 2014:377. ISSN 2229-5518,IJSER © 2014, http://www.ijser.org

[10] Kaur H, Gupta N. Localized algorithm for channel assignment in cognitive radio networks. International Journal of Engineering Research and Applications. 2015;**5**(12, (Part - 2)):171- 176, www.ijera.com. ISSN: 2248-9622

[11] Liu X, Wang Y, Chen Y, Xia J, Zhang X, Lu W, et al. A multichannel cognitive radio system design and its performance optimization. IEEE Access. 2018;**6**:12327-12335. DOI: 10.1109/ ACCESS.2018.2810061. Electronic ISSN: 2169-3536

[12] Prasannan N, Jyothish Lal G, Das R, Soman KP. Software redefined communication system. IOSR Journal of Electronics and Communication Engineering (IOSR-JECE). ISSN: 2278- 2834, ISBN: 2278-8735. www.iosrjourna ls.org. 2012;**4**(2):16-23

[13] Akeela R, Dezfouli B. Software- defined radios: Architecture, state-of- the-art, and challenges. Computer Communications, www.elsevier.com/ locate/comcom. DOI: 10.1016/j. comcom.2018.07.012

[14] Kavitha P, Meera Mohan K, Ra S, Gandhiraj R, Soman KP. Implementation of CDMA in GNU radio. Procedia Computer Science. 2015; **46**:981-988. DOI: 10.1016/j. procs.2015.01.008

[15] Anjana C, Sundaresan S, Tessy Zacharia R, Gandhiraj KPS. An experimental study on channel estimation and synchronization to reduce error rate in OFDM using GNU radio. International Conference on Information and Communication Technologies (ICICT 2014) Science Direct, Procedia computer science. 2015;**46**:1056-1063. DOI: 10.1016/j.procs.2015.01.017. www.sciencedirect.com

[16] Rebica L, Rani S, Kakkar S. Performance analysis of various modulation techniques using GNU radio. In: IJCA Proceedings on International Conference on Advances in Emerging Technology (ICAET 2016). Page no: 1–3, ICAET-Number-5 (IJCA Journal: 0975–8887); 2016

[17] Sinha D, Verma AK, Kumar S. Sample rate conversion technique for software defined radio receiver. In: 10th International Conference on Intelligent Systems and Control (ISCO); 2016. DOI: 10.1109/isco.2016.7727029

[18] Wang T, Li C. Sample rate conversion technology in software defined radio. IEEE Xplore. 2007: 1355-1358. DOI: 10.1109/CCECE.2006.277567. Print ISSN: 0840-7789

[19] Gandhiraja R, Ramb R, Somanb KP. Analog and Digital Modulation Toolkit for Software Defined Radio. Available from: www.sciencedirect.com. www.elsevier.com/locate/procedia. DOI: 10.1016/j.proeng.2012.01.9752011

[20] Lunagariya JK, Gokhruwala K, Vachhani K. Design analysis of digital modulation schemes with GNU radio. In: Second International Conference on Networks, Information & Communications 20 August 2015

[21] Elsaghier AE, Tezel NS, Altiraiki SM. Frequency shift keying scheme to implement SDR using Hackrf one. International Journal of Electronics Engineering Research. 2017;**9**(8):1147-1157. ISSN 0975-6450, © Research India Publications. Available from: http://www.ripublication.com

[22] Joshi PMP, Patil SA, Shimpi DC. Design and implementation of BPSK Audio Transmitter & Receiver Using SDR. International Journals of Advanced Research in Computer Science and Software Engineering. 2017;**7**(6):268-271. ISSN: 2277-128X

[23] Foukalas F, Khattab T, Vincent Poor H. Adaptive modulation in multi-user cognitive radio networks over fading channels. In: 2013 8th International Conference on Cognitive Radio Oriented Wireless Network. DOI: 10.4108/icst.crowncom.2013.252053

Chapter 7

Application of Random Walk Model for Timing Recovery in Modern Mobile SATCOM Systems

*Tien M. Nguyen, Hung H. Nguyen, Tom Freeze
and Andy Guillen*

Abstract

In a modern mobile satellite communication (SATCOM) system, a ground terminal receiver receives a radio frequency signal that is demodulated to generate a baseband digital signal waveform containing a self-clocking bit stream of digital data. The received baseband digital signal waveform is recovered and tracked using a timing recovery loop (TRL). The traditional TRLs use early-and-late gates, digital transition tracking, filter-and-square, and delay-and-multiply functions. In bit timing detection, the bit stream is self-clocking and the timing differential dithers about correct bit timing in the TRLs. For mobile satellite communication environments, the traditional TRLs drop lock when the loop signal-to-noise ratio (SNR) is smaller than a threshold value or the residual Doppler frequency is larger than the operating loop bandwidth. After dropping lock, the traditional TRLs experience long hang up time due to the need to reacquire the timing pulses. Recently, random walk filters (RWF) have been adapted to improve the bit clock locking stability and are applied to recover bit timing information of a digital data stream. This chapter describes random walk model for timing jitter and discusses how RWF solution can address the timing recovery challenges in mobile satellite communication environments.

Keywords: random walk filter (RWF), timing recovery loop (TRL),
Mobile communication, bit clock, phase-locked loop (PLL), square-TRL (STRL),
8PSK transmitter/receiver

1. Background and introduction

For mobile environments, current TRLs use traditional TRLs that drop lock at low signal-to-noise ratio (SNR); then, experience long hang up time while reacquiring timing pulses. A potential solution is to introduce RWF into the TRL to reduce drops, and also to shorten reacquisition time. This chapter focuses on random walk model for timing jitter and TRL using RWF [1].

A mobile SATCOM system includes a mobile transmitter, a satellite transponder and a ground terminal receiver as shown in **Figure 1**. The ground terminal receiver receives an RF signal and demodulates the RF signal to generate a baseband (BB)
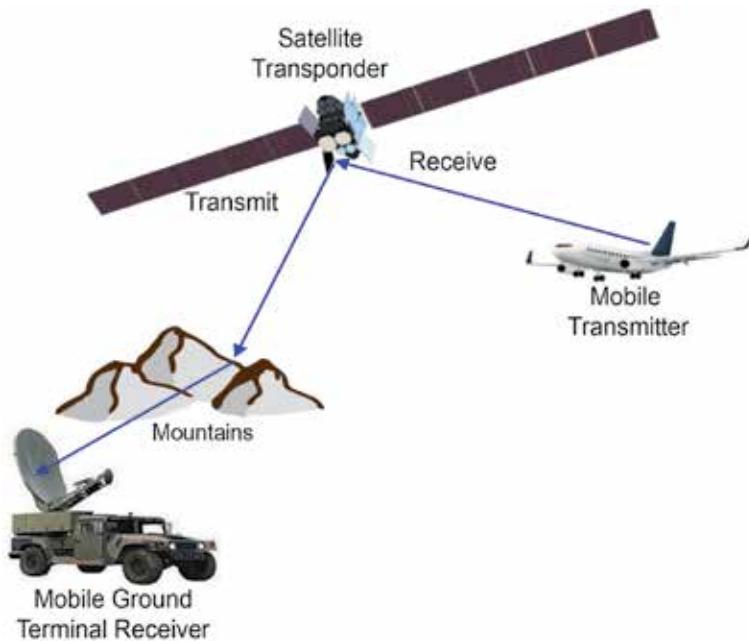
Figure 1.
Mobile satellite communication system.

signal waveform containing a self-clocking binary bit stream of digital data. The generated BB signal waveform is tracked by a TRL. The TRL tracks bit timing of the BB signal waveform and generates timing pulses. The timing pulses are then used in a receiver's data detector for sampling the BB signal waveform at the bit intervals, T_b's, for reconstructing the digital bit stream from the received BB signal waveform. The received BB signal waveform is normally distorted by channel noise causing poorly generated timing pulses and hence poor bit timing recovery leading to poor data bit detection and Bit Error Rate (BER) performance, and hence Symbol Error Rate[1] (SER) degradation. When the TRL loses track, the received binary bit stream is no longer detected. In current communication systems, the received binary bit stream is recovered using traditional TRLs that are subject to bit timing lock drop in the presence of channel noise and fading effects in mobile environment.

The traditional TRLs use Early-and-Late Gates (ELG), Digital Transition Tracking (DTT), Filter-and-Square (FaS), and Delay and-Multiply (DaM) techniques [2–11]. In bit timing detection, the binary bit stream is self-clocking and the timing differential dithers about correct bit timing in the TRLs. For mobile environments (see **Figure 1**), these TRLs drop lock when the timing Loop Signal-to-Noise Ratio (LSNR) is smaller than a threshold value or the residual Doppler frequency is larger than the operating loop bandwidth. After dropping lock, the traditional TRLs experience long hang up time due to the need to reacquire the timing pulses. RWFs have been used for decades in various applications [3–11].

In the past, RWFs have been applied to digital phase synchronization systems. RWFs have been theoretically applied to carrier phase detection where differentials between local references and transmitter carriers results in a phase correction that is unidirectional and constantly circular over 360 degrees [3–6]. Although used for decades, RWFs have not been adapted to improving the bit clock locking stability in

---

[1] For un-coded Binary Phase Shift Keying (BPSK) the SER is identical to Bit Error Rate (BER).

TRLs. This chapter discusses the RWF based on the Brownian motion process for recovering bit timing information of a binary data stream. The chapter is organized as follow:

- Section 2 presents related works, including PLL modeling using random walk theory[2], Cramér–Rao bound, Square-TRL Using Digital RWF and Advanced TRL Using RWF.

- Section 3 discusses mathematical modeling of timing jitter using random walk process and Cramér–Rao bound for timing jitter.

- Section 4 describes in detail the TRL using RWF, including software implementation and performance of RWF-TRL for 8-PSK communication system.

- Section 5 discusses the results and provides a conclusion of the chapter.

2. Related works

Section 2 describes existing publications related to RWF concept. Section 2.1 presents works related to performance modeling of PLL using random walk theory, Section 2.2 describes the Cramer-Rao Lower Bound that can be used to evaluate the best TRL performance in terms of variance, Section 2.3 discusses the operation and performance of square-TRL using digital RWF and Section 2.4 describes an advanced TRL using RWF approach.

2.1 Performance modeling of PLL using random walk theory

This subsection describes the use of random walk theory on the modeling of all digital PLL, the first-order Bang-Bang PLL and approximation of PLL phase error variance.

2.1.1 The all-digital phase-locked loop

One of the first uses of Random Walk Theory on modeling of phase-locked loop (PLL) was from [4] where the performance of an All-Digital PLL (ADPLL) was analyzed. Here a simple ADPLL for a Non-Return-to-Zero (NRZ) square wave input waveform is shown below in **Figure 2**, with an application to track the sub-carrier frequency for a satellite command system.

In this figure, the square wave signal y(t) is first low pass filtered with a bandwidth of W. The Analog-to-Digital Converter (ADC) will sample at Nyquist rate of 2 W so that there will be 2 W/f_s per period where f_s is the square wave frequency. The phase detector block consists of two elements: (1) the Transition Sampler Selector (TSS) which will output $\pm A$ at each square wave transition, since the sampling error shown as the interval τ seconds can cause the TSS output to be +A or −A depending on where τ is; and (2) an accumulator that will sum m transition samples. Following the Phase Detector, the Sign Detector will be used to add or delete one or more clock pulses, in effect shifting the sampling location of the

---

[2] The simplest model of Brownian motion is a simple symmetric random walk in one dimension, which is also referred to as random walk (see Section 3).
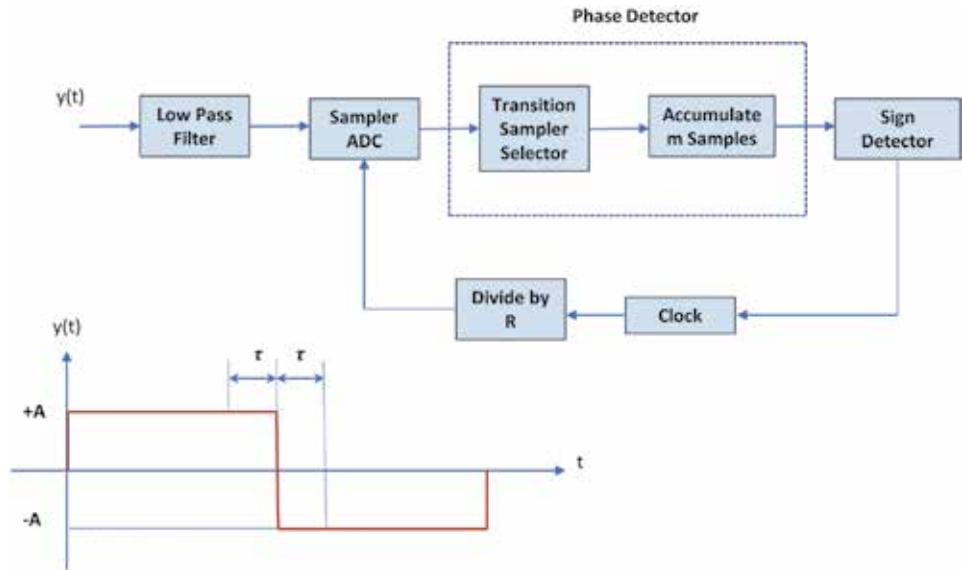
Figure 2.
Block diagram of timing loop update [4].
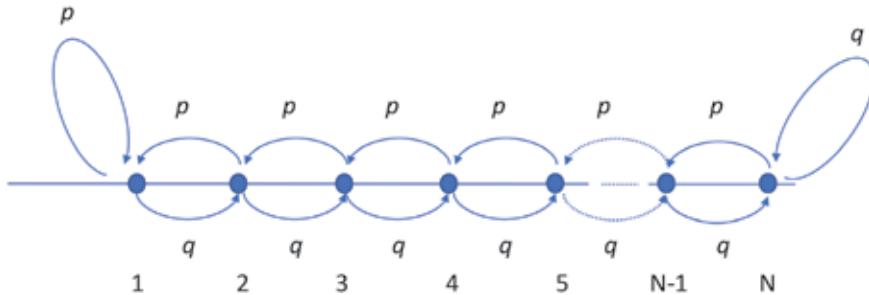
Figure 3.
Reduced state transition diagram showing a random walk [4].

square wave by a multiple of Δ defined as a fraction of the square wave period. The divide-by-R operation will convert the clock rate back to the ADC sample rate ($R = f_c/2 W$) where f_c is the clock frequency.

In the presence of white noise with power density N_0 Watt/Hz, the clock errors normalized to Δ can take on values from $-N, -(N - 1), ..., -2, -1, 1, 2, ..., (N - 1)$, N with corresponding transition probabilities. This error model forms a discrete parameter Markov chain with countable number of states, and there are $2N = 1/\Delta$ in one square wave period. The state-transition diagram can be further reduced to N total error states from 2N, by noting that the pairs of states -1 and 1, and -2 and 2, etc., contribute the same squared error. In addition, the transition probabilities are state independent since the square wave amplitude is constant for every state. This independent homogeneous finite aperiodic Markov chain is shown in **Figure 3**.

The above Markov chain is referred to as a "random walk" with the transition probabilities q and p defined in Eq. (1) and (2), respectively, and in terms of signal and noise parameters as shown in Eq. (3):

$$p = erf\left(\rho^{1/2}\right) \qquad (1)$$

$$q = 1 - p \tag{2}$$

$$\rho = mA^2/N_0 W \tag{3}$$

Where the parameter ρ is the SNR and the erf(.) is defined as in Eq. (4):

$$\text{erf}(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-t^2/2} dt \tag{4}$$

For this ADPLL, there are three major performance analysis results using the random walk model: the steady-state probabilities, timing error variance, and mean first slip time.

The steady-state probabilities P_k is defined as the probability that the PLL starting in state j is in state k when the number of updates goes to infinity. The expression for P_k is shown in Eq. (5),

$$P_k = \frac{1}{2} \frac{1 - q/p}{1 - (q/p)^k} \left(\frac{q}{p}\right)^{|k|-1}, k = -N, \ldots, -1, 1, \ldots, N \tag{5}$$

The timing error variance σ_{TE}^2 also has a closed form expression by letting α = *q/p*,

$$\sigma_{TE}^2 = \frac{\Delta^2}{4} + \frac{\Delta^2}{1 - \alpha^N} \left\{ -N(N+1)\alpha^N + 2\left[\frac{\alpha - (N+1)\alpha^{N+1}}{1 - \alpha}\right] + 2\left[\frac{\alpha^2 - \alpha^{N+2}}{(1 - \alpha)^2}\right] \right\} \tag{6}$$

It is observed from Eq. (6) that as SNR increases, the timing error σ_{TE}^2 converges to $(\Delta/2)^2$ for each value of Δ.

Since it's desirable to keep the timing error small at all times, a useful performance metrics is the Mean First Slip Time (MFST) T^{N+1} which is defined as the expected time for the timing error to exceed some magnitude (usually π or 2π) for the first time when slipping from states ± 1 to $\pm (N + 1)$. A closed form expression for this metric is shown in Eq. (7) below.
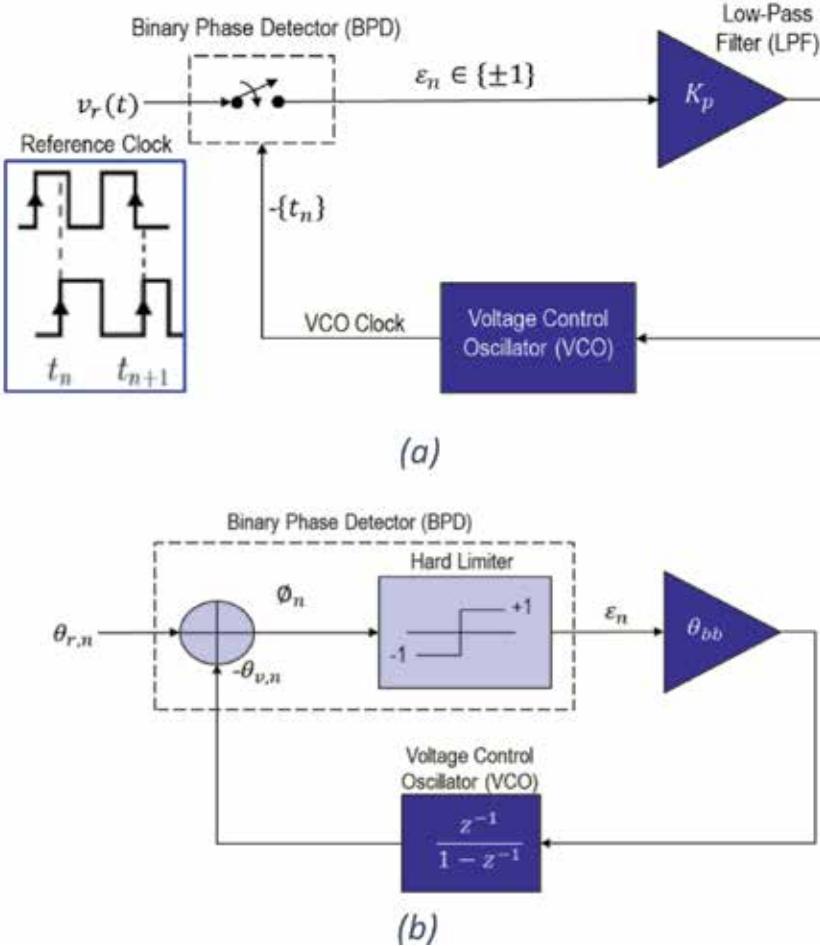
$$T^{N+1} = -\frac{N}{p - q} + \frac{\frac{1}{q} + \frac{1}{p-q}}{\frac{p}{q} - 1} \left[\left(\frac{p}{q}\right)^N - 1\right] \tag{7}$$

2.1.2 The first-order bang-bang PLL

The Bang-Bang Phase-Locked Loop (BBPLL) has been widely used for Clock and Data Recovery (CDR) in serial data links and in digital frequency synthesis [5]. A common feature in BBPLL is the Binary Phase Detector (BPD) that quantizes the phase difference between the input data and Voltage-Controlled Oscillator (VCO) clock by generating the early/late phase error information for the Loop Filter (LF). A model for the first-order bang-bang PLL is shown in **Figure 4(a)** where the BPD is represented as a sampler with input as a reference clock and sampled by a VCO. The BBPLL operation in the phase-domain can be summarized by the discrete-time model in **Figure 4(b)**.

In above figures, the behavior of the phase error $\varnothing_n$ can be described by the stochastic difference equation shown in Eq. (8) below:

$$\varnothing_{n+1} = \varnothing_n + \Delta\omega T_0 - K\left(1 + \frac{\Delta f}{f_0}\right) \text{sgn} \varnothing_n + \omega_0 \xi_n \tag{8}$$

Figure 4.
(a) Model block diagram for the first-order BBPLL [5]. (b) Equivalent phase-domain discrete time model [5].

and corresponding timing jitter is defined as in Eq. (9):

$$\Delta t_{n+1} = \begin{cases} \Delta t_n + \Delta T - K + \xi_n, \Delta t_n \geq 0 \\ \Delta t_n + \Delta T + K + \xi_n, \Delta t_n \leq 0 \end{cases} \tag{9}$$

where $\Delta T = \Delta f / f_0^2$ is the period deviation due to frequency offset[3], K is the bang-bang phase step, and ξ_n is a sequence of zero-mean independent and identically distributed (i.i.d.) random variables (RVs). This follows a random walk (RW) model starting at some origin ($n = 0$) with ξ_n being the step RV with a distribution F. This RW model is sign dependent since the next step depends on the sign of the current state, thus called Sign-Dependent Random Walk (SDRW) model. Furthermore, if we assume that $\mu_+ = \Delta T - K$ and $\mu_- = \Delta T + K$ are Gaussian RVs with variance $\sigma_{+/-}^2 = \sigma^2$ then the timing jitter process can be viewed as Sign-Dependent Gaussian Random walk (SDGRW).

PLL design questions are often centered around how much jitter is transferred from the reference clock, how much jitter is generated by the loop itself, and what is the minimum Root Mean Square (RMS) timing jitter. The first two can be answered

---

[3] The notation ΔT for frequency offset is not conventionally consistent but preserved from [5].

by examining the static timing offset error derived in [5]. A static timing offset is a significant problem in a BBPLL-based CDR circuit because it results in the incoming data no longer sampled at the center of the data eye, thus increasing the bit error rate. A closed form expression for static timing offset error is shown in Eq. (10) below.

$$\Delta t_{stat} = \Delta T + \sigma G_1\left(\frac{K - \Delta T}{\sigma}\right) + \sigma G_1\left(\frac{K + \Delta T}{\sigma}\right) \tag{10}$$

Where

$$G_k(x) = \sum_{n=1}^{\infty} g_k(n, x) \tag{11}$$

and

$$g_1(n, x) = \frac{1}{\sqrt{2\pi n}} e^{-nx^2/2} - \frac{1}{2} x \cdot erfc\left(\sqrt{\frac{n}{2}}x\right) \tag{12}$$

As shown in Eq. (10), the behavior of the static timing offset error vs. frequency offset ΔT at K = 1 can be determined by setting $\Delta T > 0$ and letting σ to be larger than 1, the combined effect of frequency offset and clock jitter causes the static timing offset to increase from its jitter-free level. But by setting $\Delta T > 0$ and letting σ be much less than 1, the static timing offset does not notably increase for a small enough σ, even with a moderate frequency offset.

The above third question can be answered by examining the variance of timing jitter error derived in [5]. A closed form expression for timing jitter error variance is shown in Eq. (13):

$$\sigma_{\Delta t}^2 = \frac{K^2}{3} + \sigma^2 + \sigma^2 G_2\left(\frac{K - \Delta T}{\sigma}\right) + \sigma^2 G_2\left(\frac{K + \Delta T}{\sigma}\right) \tag{13}$$

where

$$G_2(n, x) = \frac{1}{2}\left(nx^2 + 1\right) erfc\left(\sqrt{\frac{n}{2}}x\right) - \sqrt{\frac{n}{2\pi}} xe^{-nx^2/2} \tag{14}$$

The plot of the timing jitter error standard deviation $\sigma_{\Delta t}$ for various values of timing jitter noise σ and frequency offsets ΔT can be generated using Eq. (13) as shown in **Figure 5**. For small jitter noise, the hunting jitter dominates so RMS timing jitter error is approximately constant and approaching an asymptote. Increasing jitter noise causes RMS timing jitter error to rise because the effect of the Gaussian clock jitter and the resulting overload jitter becomes gradually apparent. For large jitter noise, overload jitter dominates, and shows a linear increase on the logarithmic scale.

Finally, an important parameter is the optimal bang-bang phase step K to achieve a minimum timing jitter variance. The expression for the optimum phase step is:

$$K_{opt}^2 = \Delta T^2 + \left(2\sigma^4 \lambda \Delta T^2\right)^{1/3} + \left(\sigma^8 / \left(24\lambda \Delta T^2\right)\right)^{1/3} \tag{15}$$

where $\lambda = 1 + \sqrt{1 - \sigma^4\left(72\Delta T^4\right)}$. The trade-off curves can be generated by substituting Eq. (15) into Eq. (13), and they can be used to optimize a BBPLL design.
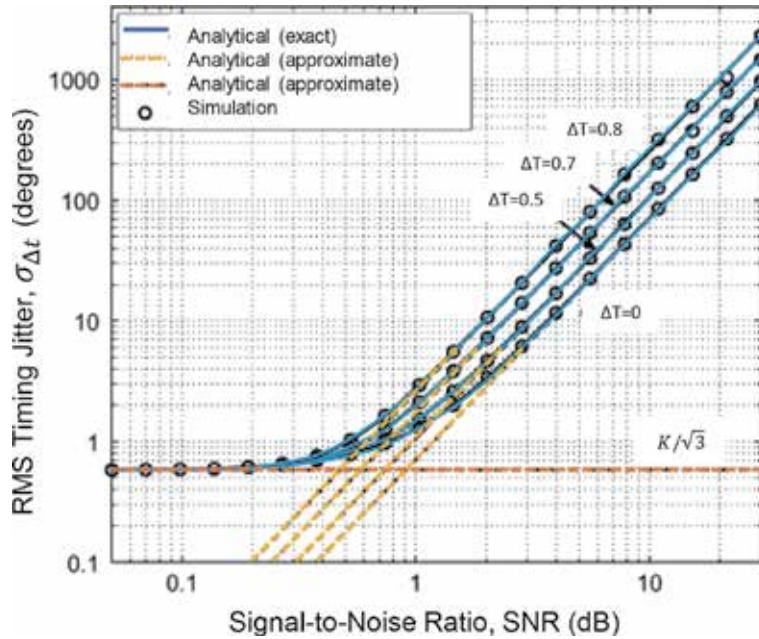
Figure 5.
RMS timing jitter error vs. SNR [5].

2.1.3 Approximation of PLL phase-error variance

Approximations for the expression of phase error variance due to nonwhite frequency noise with finite loop propagation delay have been created by Norimatsu and Ishida [7]. The PLL phase error σ_{total}^2 is expressed as a summation of the phase error variance due to phase noise, σ_{PN}^2, and the phase error variance due to shot noise σ_{SN}^2 given by [7]:

$$\sigma_{total}^2 = \sigma_{PN}^2 + \sigma_{SN}^2 \tag{16}$$

$$\sigma_{total}^2 = \int_{-\infty}^{\infty} S_{PN}(f)|1 - H(f)|^2 df + \frac{e}{M\,R\,k_s P_s} \int_{-\infty}^{\infty} |H(f)|^2 df \tag{17}$$

Where S_{PN} is the double-sided power spectral density (PSD) of the phase noise, H(f) is the closed loop transfer function of the PLL, e [A sec] is the electron charge, P_s [W] is the received optical power, R [A/W] is the photon detector responsivity, M is 4 for BPSK homodyne detection and 2 for other schemes and k_s is the optical power splitting ratio to the quadrature-arm.

2.2 Cramér–Rao bound (CRB)

In information theory and statistical analysis, the Cramer-Rao Lower Bound (CRLB) is often used to evaluate the best performance in terms of variance of any unbiased parameter estimator, whether the parameter to be estimated is random or deterministic but unknown. An unbiased estimator that achieves this lower bound is said to be statistically efficient, achieves the lowest possible mean squared error among all unbiased methods and is therefore called the minimum variance unbiased estimator (MVUE). However, there is no guarantee that an estimator exists that will achieve the CRLB. This may occur if an estimator exists, but its variance is strictly greater than the CRLB.

Given that $\hat{\boldsymbol{\theta}}(\boldsymbol{r})$ is the multiple-parameter estimator of $\boldsymbol{\theta} = [\boldsymbol{\theta_0}, \boldsymbol{\theta_1}, \dots, \boldsymbol{\theta_{M-1}}]^T$ based on the observation data vector $\boldsymbol{r} = [\boldsymbol{r_0}, \boldsymbol{r_1}, \dots, \boldsymbol{r_{N-1}}]^T$, the CRLB for this problem is given by Eq. (18):

$$\boldsymbol{CRLB} = \boldsymbol{E_{r,\theta}}\left[\left(\hat{\boldsymbol{\theta}}(\boldsymbol{r}) - \boldsymbol{\theta}\right)\left(\hat{\boldsymbol{\theta}}(\boldsymbol{r}) - \boldsymbol{\theta}\right)^T\right] \geq \boldsymbol{J_T^{-1}} \qquad (18)$$

Where $\boldsymbol{J_T^{-1}}$ is the "total" Fisher Information Matrix (FIM), defined by:

$$\boldsymbol{J_T} = \boldsymbol{E_{r,\theta}}\left\{\left[\frac{\partial}{\partial\boldsymbol{\theta}}\, \boldsymbol{ln}\, f(r,\theta)\right]\left[\frac{\partial}{\partial\boldsymbol{\theta}}\, \boldsymbol{ln}\, f(r,\theta)\right]^T\right\} \qquad (19)$$

Where $f(\boldsymbol{r},\boldsymbol{\theta})$ is the joint probability density function (pdf) of $\boldsymbol{r}$ **and** $\boldsymbol{\theta}$. The estimation error variance of each estimate of $\boldsymbol{\theta_i}$ can be taken from the diagonal term of the $\boldsymbol{J_T^{-1}}$ matrix. Since $f(\boldsymbol{r},\boldsymbol{\theta}) = \boldsymbol{f_{r|\theta}}(r|\theta)\boldsymbol{f_\theta}(\theta)$ where $\boldsymbol{f_{r|\theta}}(r|\theta)$ is the conditional pdf of $\boldsymbol{r}$ given $\boldsymbol{\theta}$, and $\boldsymbol{f_\theta}(\theta)$ is the a-priori pdf of $\boldsymbol{\theta}$. So the total FIM can be separated into two distinct parts representing deterministic and random parts for each type of estimator is shown in Eq. (20):

$$\boldsymbol{J_T} = \boldsymbol{E_\theta[J_\theta]} + \boldsymbol{J_{ap}} \qquad (20)$$

Where the deterministic FIM part is:

$$\boldsymbol{J_\theta} = \boldsymbol{E_\theta}\left\{\left[\frac{\partial}{\partial\boldsymbol{\theta}}\, \boldsymbol{ln}\, \boldsymbol{f_{r|\theta}}(r|\theta)\right]\left[\frac{\partial}{\partial\boldsymbol{\theta}}\, \boldsymbol{ln}\, \boldsymbol{f_{r|\theta}}(r|\theta)\right]^T\right\} \qquad (21)$$

and the a-priori FIM for the random parameter estimator is:

$$\boldsymbol{J_{ap}} = \boldsymbol{E_\theta}\left\{\left[\frac{\partial}{\partial\boldsymbol{\theta}}\, \boldsymbol{ln}\, \boldsymbol{f_\theta}(\theta)\right]\left[\frac{\partial}{\partial\boldsymbol{\theta}}\, \boldsymbol{ln}\, \boldsymbol{f_\theta}(\theta)\right]^T\right\} \qquad (22)$$

For each problem, the appropriate FIM will need to be calculated accordingly. The inversion of the composite FIM will result in the CRLB.

2.3 Square-TRL using digital RWF

Digital filter with Square-TRL (DF-STRL) has been proposed and discussed in [6]. Due to its simplicity and hang-up-free filtering features, DF-STRL has been used by satellite ground terminals for Pulse Amplitude Modulation (PAM), Quadrature Amplitude Modulation (QAM) and Phase Shift-Keying (PSK) waveforms operating at high data rates. A typical DF-STRL architecture is shown in **Figure 6**, where the integer N is selected to be 4.
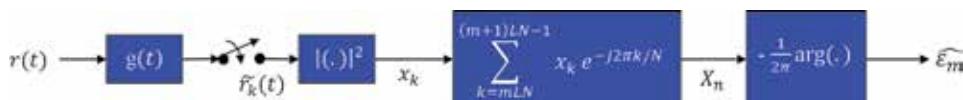
Figure 6.
Digital filter and square-TRL (DF-STRL) [6].

If one assumes that the mean of the timing estimate $\hat{\varepsilon}_m$ is zero, the timing jitter (or timing variance of the timing estimate), σ_τ^2, is defined as:

$$\sigma_\tau^2 = E\left\{(\hat{\varepsilon}_m)^2\right\} = \frac{1}{(2\pi)^2} E\left\{(\arg(X_n))^2\right\} = \frac{1}{(2\pi)^2} E\left\{\left(\frac{\operatorname{Im}(X_n)}{\operatorname{Re}(X_n)}\right)^2\right\} \qquad (23)$$

Reference 6 shows that the timing jitter, σ_τ^2, for DF-STRL consists of three components, namely

- Timing jitter generated by (Signal x Signal)–This term is referred to as self-noise term: σ_{SxS}^2

- Timing jitter generated by (Signal x Noise)–This term is referred to as Squaring Loss (SL) term: σ_{SxN}^2

- Timing jitter generated by (Noise x Noise): σ_{NxN}^2

Mathematically, the timing jitter, σ_τ^2, for DF-STRL can be expressed as:

$$\sigma_\tau^2 = \sigma_{SxS}^2 + \sigma_{NxN}^2 + \sigma_{SxN}^2 \qquad (24)$$

Analysis shown in [6] showed that the timing estimate $\hat{\varepsilon}_m$ is un-biased, which means that the average $\hat{\varepsilon}_m$ coincides with ε_m. The estimation accuracy depends on several loop parameters, including Symbol SNR, observation interval L_0 and the pulse shaping filter g(t). The timing jitter in Eq. (24) can be put in the form ([11], Section 7.6.2):

$$\sigma_\tau^2 = \frac{T}{L_0}\left[K_{SxS} + K_{SxN}\left(\frac{E_S}{N_0}\right)^{-1} + K_{NxN}\left(\frac{E_S}{N_0}\right)^{-2}\right] \qquad (25)$$

where T is the symbol duration, $\frac{E_S}{N_0}$ is the symbol SNR, and K_{SxS}, K_{SxN} and K_{NxN} are the coefficients depending on the symbol alphabet and the pulse shaping filter g(t). The pulse shaping filter g(t) is usually Root-Raised-Cosine filter with roll-off α.

2.4 Advanced TRL using RWF

Recently, a small team of Aerospace engineers developed an innovative approach to recover the timing information from a received binary data stream [1]. The proposed approach uses:

- A "RWF counter" for counting early, nominal and late arrivals of data transition pulses of an input binary data stream

- The output of the RWF provides magnitude counts that are compared to a Threshold Value (TV) that when exceeded by the magnitude counts results in a delay adjustment of the generated Adjusted Timing Pulses (ATPs)

- The ATPs are eventually catching up with the actual timing clocks

- When ATPs caught up, no delay adjustment will be made allowing ATPs to synchronize with the actual bit timing for maintaining bit timing lock
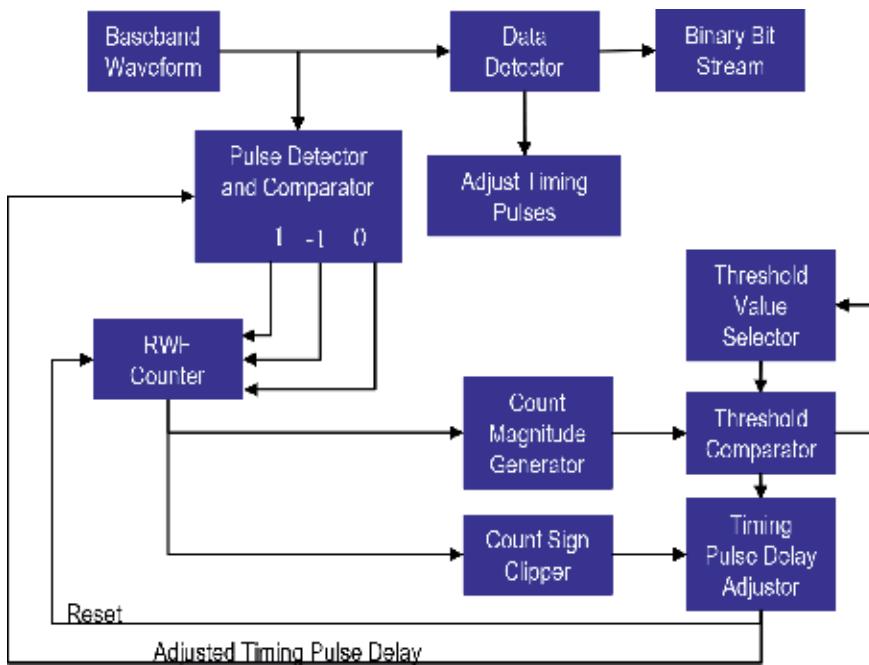
Figure 7.
RWF TRL [1].

- The ATPs are used by a data detector for reliable data detection and reconstruction of the binary bit stream.

With the proposed RWF approach, the threshold value "TV" can be adaptively adjusted for reducing drop lock rates in the presence of changing channel environments. **Figure 7** illustrates the proposed RWF approach for timing recovery. Section 4 discusses in depth the proposed RWF implementation approach and presents simulation results for 8PSK Modulator-Demodulator (MODEM).

3. Mathematical modeling of timing jitter using random walk process

The Wiener process was introduced as a mathematical model of Brownian motion describing a random, but continuous motion of a particle, subjected to the influence of a large number of chaotically moving molecules of the liquid. The simplest model of Brownian motion is a simple symmetric random walk in one dimension (aka random walk) [12]. It also has been shown that Brownian motion with zero drift is the limiting case of random walk [13]. This section describes the mathematical random time walk model for charactering the timing jitter in an Additive White Gaussian Noisy channel and the corresponding Cramér–Rao bound for timing jitter.

3.1 Mathematical modeling of random time walk

References [3, 5] discuss a Random Time Walk (RTW) model, which can be captured in the following **Figure 8**. The RTW model includes a binary data stream $\{a_k\}$ that is independent identically distributed data symbols (-1 or 1), a timing

Figure 8.
Random time walk model [3].

offsets τ timing, a channel impulse response h(t), an Additive Gaussian White Noise (AWGN) source and channel output y(t).

The RTW model used for the modeling of timing jitter and performance analysis shown in **Figure 8** [3], can be expressed mathematically as:

$$y(t) = \sum_{l=0}^{N-1} a_l h(t - lT - \tau_l) + n(t) \tag{26}$$

where T is the bit period, $a_l \epsilon \{\pm 1\}$ are the N i.i.d. and equally likely data symbols, $h(t)$ is the channel impulse response, $n(t)$ is additive white Gaussian noise, and τ_l is the random timing offset for the l-th symbol. This timing offset model can also be expressed as:

$$\tau_{k+1} = \tau_0 + (k+1)\Delta T + \sum_{l=0}^{k} w_l \tag{27}$$

where τ_0 is the initial timing offset and zero-mean Gaussian random variable with variance $\sigma_{\tau_0}^2$, ΔT is the frequency offset parameter zero-mean Gaussian random variable with variance $\sigma_{\Delta T}^2$, $\{w_l\}$ characterizes a random walk and are i.i.d. zero-mean Gaussian random variables of variance σ_w^2 which determines the severity of the random walk.

3.2 Cramér–Rao bound (CRB) for timing jitter

The RTW model described in Section 3.1 will be used for the CRLB derivation. To eliminate out-of-band noise at the receiver, the received waveform $y(t)$ is filtered by a front-end filter with impulse response $f(t)$ to get the waveform $r(t)$, which will be sampled at instant kT to get baud-rate sample $\{r_k\}$ given by:

$$r_k = \sum_{l=0}^{N-1} a_l h_1(kT - lT - \tau_l) + n_k = x_k + n_k \tag{28}$$

Where $h_1(t) = h(t) * f(t)$, $x = [x_0, x_1, \ldots, x_{N-1}]^T$ is the vector of signal component of r_k, and n_k are zero-mean i.i.d. normal random variables with variance σ^2. In this timing jitter models, the random parameters to be estimated are $a = [a_0, a_1, \ldots, a_{N-1}]^T$ is the vector of transmitted symbols, and $\tau = [\tau_0, \tau_1, \ldots, \tau_{N-1}]^T$ is the vector of timing offsets.

For the CRLB computation, the parameter vector $\theta = [\tau^T \, a^T]^T$ and $\hat{\theta} = [\hat{\tau}^T \hat{a}^T]^T$. From Section 2.2,

$$J_T = \begin{bmatrix} J_T^\tau & 0 \\ 0 & J_T^a \end{bmatrix} \tag{29}$$

represent the total FIM so that the joint estimation is decoupled, and to get the CRLB on timing estimation it is sufficient to evaluate and invert J_T^τ which is given by

$$
J_T^\tau = \frac{1}{\sigma_w^2}
\begin{bmatrix}
\beta_2 & -1 & 0 & \dots & 0 \\
-1 & \lambda & -1 & \ddots & \vdots \\
0 & \ddots & \ddots & \ddots & 0 \\
\ddots & \ddots & -1 & \lambda & -1 \\
0 & \ddots & 0 & -1 & \lambda - 1
\end{bmatrix}
\tag{30}
$$

Where $\lambda = 2 + C\frac{\sigma_w^2}{\sigma^2 T}$, $\beta_2 = \beta_1 + C\frac{\sigma_w^2}{\sigma^2 T}$, and $\beta_1 = \left(\frac{N^2 - N - 1}{N^2} + \frac{\sigma_w^2}{\sigma_{\tau_0}^2} + \frac{\sigma_w^2}{N^2 \sigma_{\Delta T}^2}\right)$, and $C = \left(\frac{2\pi^2}{3} - 1\right)$. The a-priori information parameters $\sigma_{\tau_0}^2$ and $\sigma_{\Delta T}^2$ represent the uncertainty of the initial timing offset and frequency offset, being zero would mean perfect knowledge of these two quantities which is not practical, and the estimation problem is simply that of estimating a random walk. Inverting (30) to get the CRLB for each τ_i leads to

$$
\frac{E\left[(\hat{\tau}_i(r) - \tau_i)^2\right]}{T^2} \geq h \cdot f(i)
\tag{31}
$$

where

$$
h = \frac{\sigma_\varepsilon^2}{T^2} \frac{\eta}{\eta^2 - 1}
\tag{32}
$$

is the stead state value of the CRLB,

$$
f(i) = \tanh\left(\left(N + \frac{1}{2}\right)\ln\eta\right)\left[1 - \frac{\sinh\left((N - 2i - \frac{3}{2})\ln\eta\right)}{\sinh\left((N + \frac{1}{2})\ln\eta\right)}\right]
\tag{33}
$$

and $\eta = \frac{\lambda + \sqrt{\lambda^2 - 4}}{2}$.

The CRLB derived applies to timing recovery systems in general. In practice, timing recovery systems usually involve phase locked loops. A traditional PLL is typically used for timing recovery, and the receiver employs a timing-error detector (TED) to arrive at timing-error estimates. For simplicity, a first-order PLL is employed which updates its estimate according to

$$
\hat{\tau}_{k+1} = \hat{\tau}_k + \alpha \hat{e}_k
\tag{34}
$$

where α is the PLL gain, and $\hat{e}_k$ is the detector estimate of the estimation error. Instead of feeding the TED decisions about the received symbols, if we allow it to have access to the actual transmitted symbols, then we have a trained PLL. The performance of trained PLL gives a heuristic lower bound for the performance of receiver structures that use the PLL for timing recovery.

In **Figure 9**, the steady state CRB and the performance of the trained PLL are plotted for the following system parameters: $\frac{\sigma_\varepsilon}{T} = 0.5\%$, block length $N = 500$, and the PLL performance being averaged over 1000 trials [8]. As seen in the figure, the performance of the PLL is a strong function of the gain parameter α, and therefore it has to be optimized for each SNR. The PLL error variance is plotted for various
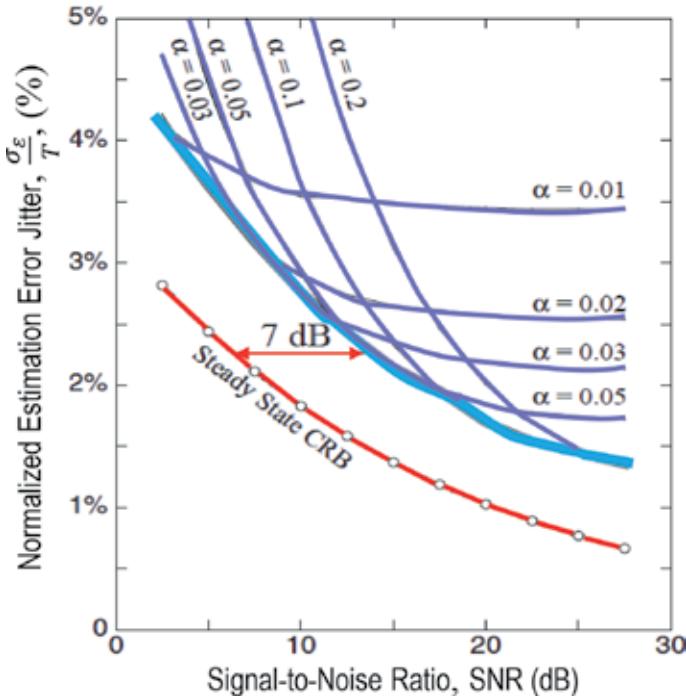
Figure 9.
Trained PLL and the CRLB [8].

values of α. Taking the minimum of the error variance over all α gives us the best performance we can expect using the trained PLL. We see that the trained PLL is about 7 dB away from the CRB. This gap of 7 dB has to be put in perspective by the fact that the CRB is not attainable in this case. For the CRB to be attainable, the a-posteriori density $f_{r|\theta}(r|\theta)$ needs to be Gaussian, which is not the case here.

4. Timing recovery using RWF

The most commonly used TRLs in existing wireless and satellite communications systems are DF-STRL, Early-and-Late Gate, Digital Transition Tracking, and Delay-and-Multiply TRLs [6, 9–11]. For mobile environments, these TRLs drop lock when the loop signal-to-noise ratio (SNR) is smaller than the threshold value SNR or the residual Doppler frequency is larger than the operating loop bandwidth. When "dropped lock", these timing loops experience a long "hang-up" time due to reacquisition and locking behavior of the timing clock. As discussed in Section 2.3 on the related works for DF-STRL, a form of digital RWF was proposed for use with square TRL. This DF-STRL suffers squaring loss caused by squaring mechanism used by the loop. This section explores the advanced TRL using RWF.

4.1 RWF concept for timing recovery

The RWF-TRL concept derived from the following principles:

- Principle 1: Input binary signal waveform is compared in time with locally generated and suitably delayed timing pulses to produce lead/lag signals to increment/decrement a timing counter.

- Principles 2: When timing counter output exceeds a positive or negative threshold, the delay for timing pulses is adjusted accordingly.

- Principle 3: Without timing errors, an appropriately selected threshold value allows the lead and lag signals to cancel out in time, thus retaining the correct timing pulses.

The block diagram, shown in **Figure 10**, describes the above principles pictorially. The received binary data is a square wave with amplitude varying between +1 and −1 with time duration T in second, which is inversely proportional to the communication data bit rate R in bit per second.

4.2 Software implementation of RWF TRL

The major Software Blocks (SWB) required for implementing the above three principles are:

- SWB 1: Pulse detection and comparison block converts input baseband digital signal waveform to data transition pulses and compares them to timing pulses to generate lead/lag signals.

- SWB 2: RWF counter block generates a running count that is incremented/decremented by the lead/lag signals.

- SWB 3: Threshold comparison block generates exceedance signal when magnitude of running count is greater than a selected threshold.

- SWB 4: Timing pulse delay adjustment block uses sign of running count to adjust timing pulses when triggered by exceedance signal.

Figure 11 presents a software architecture for implementation of the three principles described in Section 4.1. The threshold value is derived based on a threshold selection process that depends on the channel environment. Either a training sequence or a priori knowledge of the channel propagation conditions may
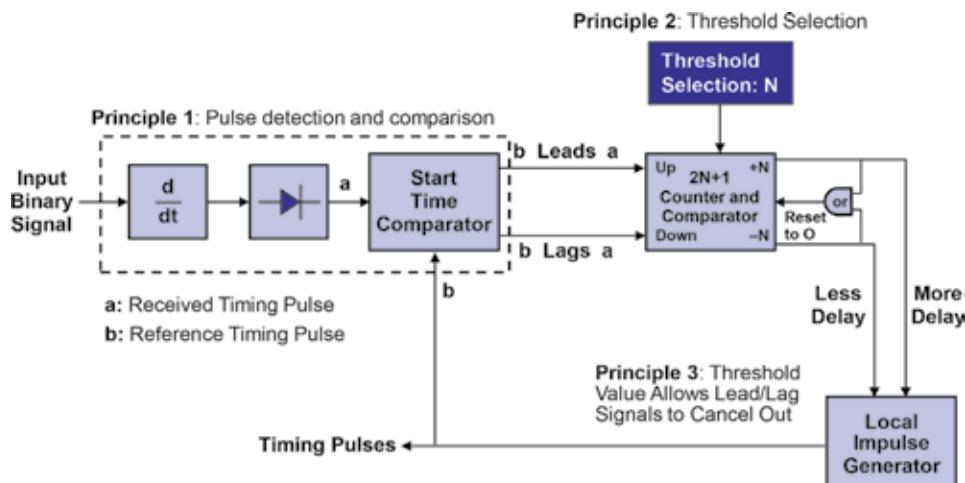
Figure 10.
RWF concept for timing recovery.

be needed to set the threshold. Alternatively, an algorithm can be provided to adaptively select a threshold value to compensate for channel impairments.

Figure 12 describes the "Pulse Detection and Comparison Block", or SWB 1. This SWB1 consists of the following functions:

- Converts input baseband digital signal waveform to data transition pulses delayed by half of the search window size, W.

- Generates timing pulses that are delayed by the adjusted timing pulse delay that is input to the block.

- Counts number of data transition pulses that are within W from each successive timing pulse and keep track of the delay of the first data transition pulse.

- Calculates lead/lag time of the first data transition pulse as the delay time minus half of the search window size, and lead/lag signal as the sign of the lead/lag time.

- Outputs a non-zero lead/lag signal if there is one, and only one, data transition pulse within the search window, thus eliminating much of the noise-caused ambiguity.
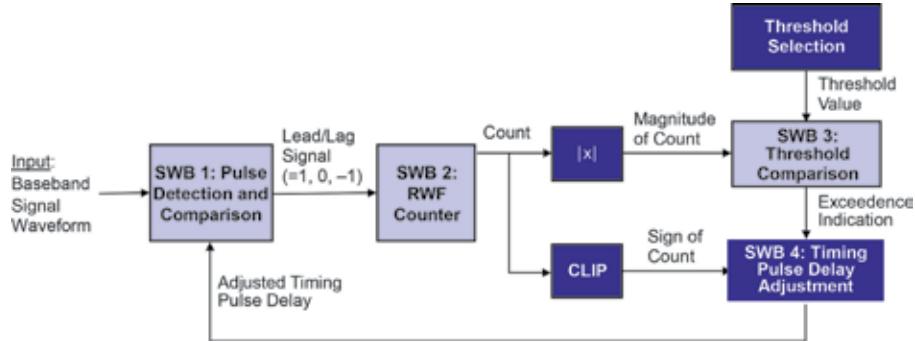
Figure 11.
Software architecture of advanced TRL using RWF.
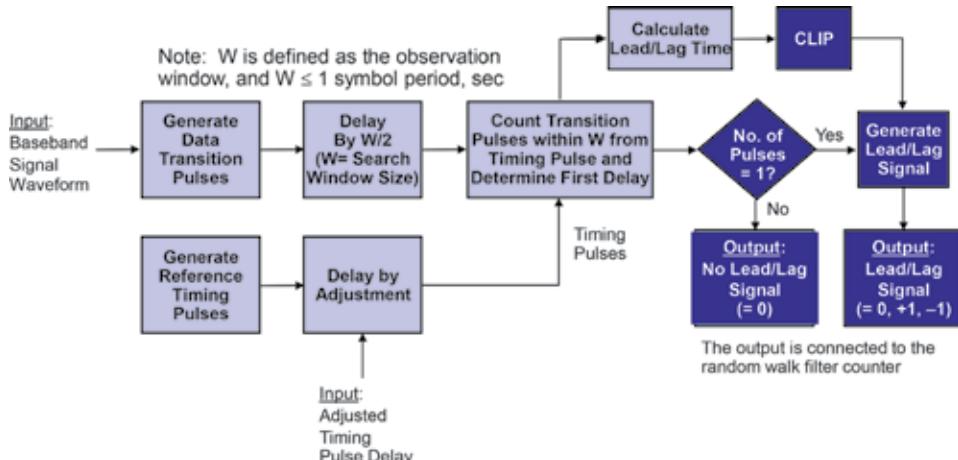
Figure 12.
SWB 1 implementation of pulse detection and comparison.

114

4.3 Performance of RWF TRL

Signal Processing Work (SPW) implementation of the proposed RWF TRL for 8-PSK modem is described in **Figure 13**. The proposed implementation includes the following blocks:

- 8PSK transmitter block generates signals modulated by sine and cosine waveforms.

- AWGN generator block adds Gaussian random noise to the 8PSK signals.

- Demodulator with ideal phase tracking converts the 8PSK signals to baseband.

- Combination of low pass filter (LPF) and hard-limiter generates a binary data stream from the 8PSK signals.

- RWF TRL block generates clock signals (timing pulses).

- Sum dump and hold block uses the clock signals to integrate the baseband 8PSK signals.

- 8PSK detector puts the integrated signals through 8PSK slicer to reconstruct the data stream.

- SER estimator compares the original data stream with the reconstructed data stream to detect symbol errors.

SPW-simulated SER curve with imperfect timing recovery is shown in **Figure 14**. This SER curve was obtained under the following operational conditions, which is not optimized in terms of search window and threshold value used by the counter:

- Back-to-back Modem with no transponder or amplifier in between, i.e., no Amplitude Modulation-to-Amplitude Modulation (AM-AM) and Amplitude Modulation-to-Phase Modulation (AM-PM) distortions.

- Carrier frequency: 5000 Hz
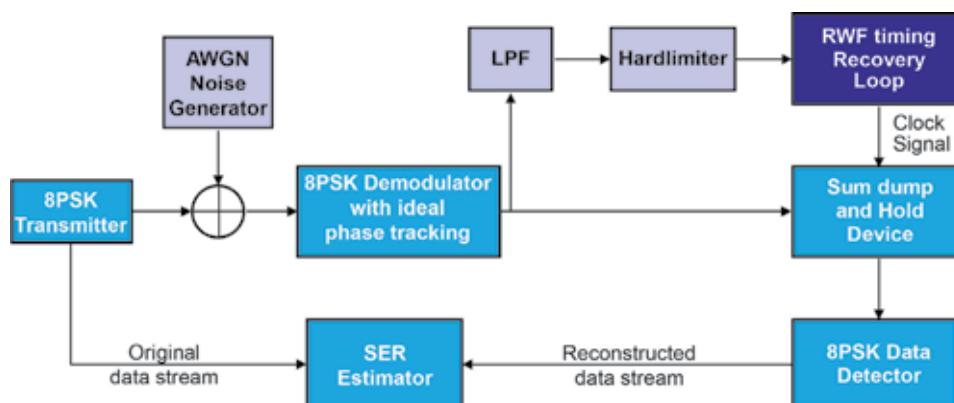
Figure 13.
Simulation set-up for 8PSK modem with RWF-TRL.
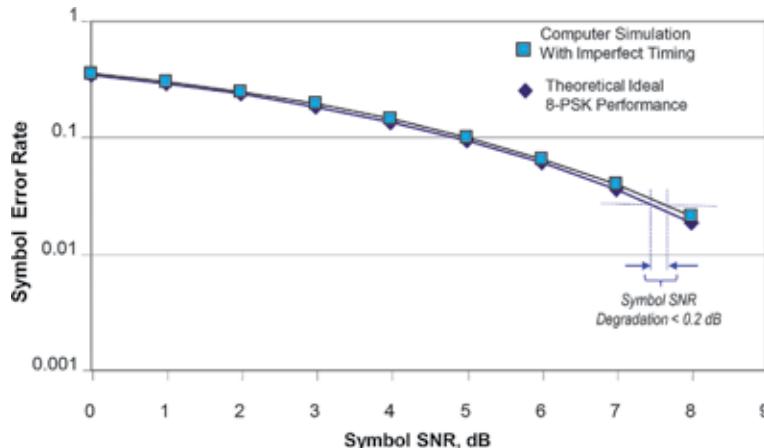
Figure 14.
8-PSK symbol error rate simulation with RWF TRL.

- Symbol rate: 1000 sps

- Number of samples per symbol: 50

- Search window size: 10 samples

- Threshold for counter: 10 counts

For symbol SNR greater than 2 dB, the theoretical SER for 8PSK can be accurately estimated from the following Eq. [9–10]:

$$SER \approx 2Q\left\{ \sqrt{2.\gamma_s}\, sin\left(\frac{\pi}{8}\right) \right\} \tag{35}$$

where is γ_s the symbol SNR and $Q(x)$ is the Q-function given by:

$$Q(x) = \frac{1}{\sqrt{2\pi}} \int_x^\infty e^{\left(-t^2/2\right)} dt = 1 - \text{erf}(x) \tag{36}$$

As shown in **Figure 14**, the simulated SER curve for 8PSK with perfect phase tracking and timing recovery coincides with the theoretical SER curve for 8PSK. Simulated SER curve with imperfect timing (but perfect phase tracking) shows a symbol SNR degradation of about 0.2 dB from the theoretical curve.

5. Discussion and conclusion

Current trends in mobile communications will be impacted from operational environment, many near-by users and spectrum sharing, expanding amounts of data (such as streaming video), and increased mobility of receivers. As such, there is higher potential to drop lock, and greater impact resulting from traditional recovery times. A proposed solution is to introduce RWF into the TRL to reduce drops, and also shorten reacquisition time. Based on the RWF-TRL simulation results for 8PSK comparing to standard TRLs for similar 8PSK modem, we believe that the proposed loop has the following advantages:

- Using adaptable threshold values of the RWF timing loop that is trained for specified operational environments, the "dropped lock" rate is expected to be less

- Faster acquisition time due to the adaptable threshold value setting of the RWF timing loop

- Can be made to be adaptive to any mobile environment using deep learning and artificial intelligent technology.

- All digital and simple to implement

- Expandable to accommodate a wide range of modulation schemes, namely, BPSK, QPSK, 16-QAM, 64-QAM, and possible application to continuous phase modulation, e.g., GMSK.

- The authors searched for similar SER results for 8PSK Modem to compare with the simulation results presented in Section 4.3, but currently, there are no TRL using RWF available for performance comparison.

As described in Section 4, there are several features associated with RWF approach, including:

- Operate at low sampling rate, which means less power consumption.

- Provide excellent SER performance

- Robust to mobile operational environment due to adaptable feature

- Accommodate a wide range of modulation schemes

The potential markets for the proposed RWF filter approach include:

- Mobile satellite industry: Mobile satellite terminals, satellite decoders, etc.

- Wireless communication industry: Cellular and mobile phone industry.

Acknowledgements

Conflict of interest

Using the data reported in [1], which was performed under The Aerospace IRAD funding in early 2000's, the authors prepared Sections 2.4 and 4 of this chapter. However, the preparation of this chapter was not funded by The Aerospace Corporation, and it was done by the authors using their own time and thus it does not

represent The Aerospace Corporation's view on the use of RWF-TRL for future mobile satellite systems.

Notes/Thanks/Other declarations

The first author wishes to thank his wife, Thu-Hang Nguyen, for her enormous patience and boundless support during the preparation of this chapter.

Author details

Tien M. Nguyen*, Hung H. Nguyen, Tom Freeze and Andy Guillen
The Aerospace Corporation, El Segundo, California, USA

*Address all correspondence to: tien.m.nguyen@aero.org

IntechOpen

References

[1] Nguyen Tien M, Yoh James, Mathur Ashok, Goo Gary. Random Walk Timing Recovery Loop, U.S. Patent, Patent No. US 7.469,026 B2, 2008

[2] Sun J. Synchronization for capacity-approaching coded communication systems [Ph.D. dissertation]. Morgantown, West Virginia: West Virginia University; 2004

[3] Nayak AR, Barry JR, Feyh G, McLaughlin SW. Timing recovery with frequency offset and random walk: Cramér–Rao bound and a phase-locked loop postprocessor. IEEE Transactions on Communications. 2006;**54**:2004-2013

[4] Holmes J. Performance of a first-order transition sampling digital phase-locked loop using random-walk models. IEEE Transactions on Communications. 1972;**20**:119-131

[5] Tertinek S, Gleeson JP, Feely O. Statistical analysis of first-order bang-bang phase-locked loops using sign-dependent-random-walk theory. IEEE Transaction on Circuits and Systems. 2010;**57**:2367-2380

[6] Oerder M, Meyr H. Digital filter and square timing recovery. IEEE Transactions on Communications. 1988;**36**:605-612

[7] Norimatsu S, Ishida O. Impact of flicker noise and random-walk noise on a phase-locked loop with finite propagation delay. IEEE Transactions on Communications. 1994;**12**:86-95

[8] Nayak AR, Barry JR, McLaughlin SW. Lower Bounds for the Performance of Iterative Timing Recovery at Low SNR [Technical Report]. Atlanta: Georgia Institute of Technology; 2002. Available from: http://barry.ece.gatech.edu/pubs/conference/mtns02.pdf

[9] Holmes JK. Coherent Spread Spectrum Systems. 2nd ed. New York: Wiley and Sons; 1982. 1132p

[10] Lindsey WC, Simon MK. Telecommunication Systems Engineering. New Jersey: Prentice-Hall; 1973

[11] Meyr H, Moeneclaey M, Fechtel SA. Digital Communication Receivers: Synchronization, Channel Estimation, and Signal Processing. New York: John Wiley and Sons; 1998

[12] Szabados T. An Elementary Introduction to the Wiener Process and Stochastic Integrals [Technical Report]. Budapest, Hungary: Technical University of Budapest; 1991. Mathematics Subject Classification. Primary 60J65, 60H05. Secondary 60J15, 60F15. Available from: https://arxiv.org/pdf/1008.1510.pdf

[13] Lyuu Y-D. Brownian Motion as Limit of Random Walk [Technical Report]. Taipei, Taiwan: National Taiwan University; 2014. pp. 519-597. Available from: https://www.csie.ntu.edu.tw/~lyuu/finance1/2014/20140423.pdf

Design Principles of 5G NR RoF-Based Fiber-Wireless Access Network

Mikhail E. Belkin, Tatiana N. Bakhvalova
and Alexander S. Sigov

Abstract

For today, much attention in the upcoming 5G New Radio (NR) mobile networks is paid to radically expanding the available spectral bands up to millimeter wavelengths (MMW). Following this tendency, currently, the local telecommunication commissions of various countries are proposing and harmonizing the plans of frequency allocation in MMW band, which will be reviewed this year at the World Radio Conference (WRC-2019). Another milestone of great importance is the development of access networks. Here, well-known radio-over-fiber (RoF) technology is considered as the most promising approach, which is implemented based on fiber-wireless (FiWi) architecture. Elaborating the direction, in this chapter we review the worldwide progress of RoF-architected 5G NR access networks and highlight our last simulation results on design and optimization of millimeter-photonic-based FiWi interface. All schemes are modeled using VPIphotonics Design Suite software tool. In the result of simulation experiments, optimal design principles of optical distribution network (ODN), fiber-wireless interface (FWI), and fiber-wireless fronthaul network (FWFN) as a whole have been proposed, described, and validated.

Keywords: 5G NR mobile communication system, access network, RoF technology, millimeter-wave fiber-wireless architecture, computer-aided design

1. Introduction

Within the recent decades, an explosion of researches and developments referring to the next-generation communication networks known as 5G New Radio (NR) has been observed [1–5]. Based on 4G long-term evolution (LTE) progress [6], 5G NR is in principle a novel stage of unprecedented technological innovation with ubiquitous speed connectivity. As a result, it is expected that 5G NR will radically transform a number of industries and will provide direct, superspeed connections between any users, sensors, and devices.

At the present time, several reviews to analyze significant changes in the 5G NR approaches as compared to the existing 4G LTE networks have been published [7, 8] denoting a series of milestones. Among them, much attention is paid to radically expanding the available spectral bands up to millimeter wavelengths (MMW). Following this tendency, currently, the local telecommunication
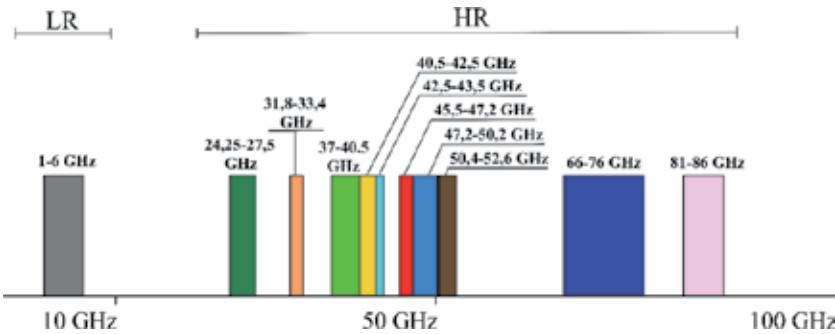
Figure 1.
Planned 5G NR spectrum allocations [10].

commissions of various countries are proposing and harmonizing the plans of frequency allocation in microwave band that has to be coexistent with 4G LTE and in newer MMW band, which will be reviewed this year at the World Radio Conference (WRC-2019) [9]. However, **Figure 1** shows a preliminary frequency plan [10] sharing two separate sub-bands: the so-called low range (LR) inside 1–6 GHz and the high range (HR) inside 24–86 GHz.

Another milestone of great importance is the development of access networks. In this direction, the well-known radio-over-fiber (RoF) technology [11–13] is considered as the most promising approach, which is implemented based on fiber-wireless (FiWi) architecture.

Following them, recently we contributed some works referred to design microwave-photonic-based MMW FiWi interface [14–24]. Elaborating the direction, in this chapter we review the worldwide progress of RoF-architected 5G NR access networks and highlight our last simulation results on design and optimization of photonic-based FiWi interface. In this way, the rest of the chapter is organized as follows. Section 2 reviews the distinctive features of access networks for 5G NR mobile communication systems including small cell scenario, RoF concept, and microwave-photonic-based approach to construct the network equipment. A specific example illustrating a RoF-based small cell scenario in 5G NR network is also included. In addition, Section 3 presents the results of our recent investigations to design optimally a fiber-wireless fronthaul network (FWFN) including an optical distribution network (ODN) and MMW fiber-wireless interface (FWI). All schemes are modeled using off-the-shelf VPIphotonics Design Suite software tool. Finally, Section 4 concludes the chapter.

2. The distinctive features of access networks for 5G NR mobile communication systems

Generally, low transmission loss and broad bandwidth characteristics of optical fibers, possibility of wavelength division multiplexing, and low sensitivity to electromagnetic interference of optical fiber-based transmission systems allow the introduction of novel concepts into distribution and processing of the digital signals being transmitted over a communication network. One of the most important examples of introducing radically new approaches is the upcoming mobile network 5G NR. According to the Introduction, a number of new principles are introduced in 5G NR network design. Three of the most suitable for access networks will be discussed below.

2.1 Small cell scenario

In the last decade, the problem of developing and optimizing the architecture of the fifth-generation communication networks and developing equipment for their implementation has received the closest attention of the global telecom community [6]. It is predicted that the implementation of these next-generation networks will provide unprecedented amounts of data and services for mobile and fixed users, which can be called both an evolution and a revolution in mobile cell technologies [1]. They are architectural in nature—for example, moving some decision-making to the devices themselves (device-centric architectures and smart devices)—or most networks are hardware-oriented. Besides, continuously increasing requirements for broadband services and capacity of communication links by enhancing the data transfer rate in all sections of the cellular network led to the shift of the operating frequency to millimeter-wave band, with a total cell capacity of several gigabits per second. One of the keys of them is ultra-densification of service areas and users. The data obtained from the analysis of a large number of publications, on the quantitative parametric comparison for mobile communication networks of the available fourth and incoming fifth generations, are presented in **Table 1**.

To ensure so sharp explosion of the key parameters, a significant complication of the standard cellular network structure is required. Thus, according to the generally accepted opinion, the ambitious goals for the development of fifth-generation wireless networks can be achieved by solving two advanced global tasks: architectural one, associated with the small cell scenario for access network, and the technological one, associated with the introduction of microwave-photonic (MWP) approach to the design of the network equipment. The further is connected with the introduction of the fronthauls based on FiWi architecture. The latter is especially important for interface network units, both between the fiber-optic backhaul network and fronthaul networks, whose task is to transfer high-speed data stream to millimeter-wave carriers, and between wired and wireless sections of the access network, in which the signals of the optical and millimeter-wave bands should be cost-efficiently converted. As an example of implementing small cell scenario concept, an advanced skeleton diagram of 5G NR network using a common central station (CS), fiber backhauls, and FiWi fronthauls is shown in **Figure 2**.

2.2 Radio-over-fiber concept

Analyzing the parameters of **Table 1**, one can make the following clear conclusion. More than an order of magnitude increased requirements to throughput make fiber-optic communication system as a leading technology not only for transport networks but also for next-generation access networks. However, the important

Parameter	4G LTE	5G NR
Connection density (per km^2)	Less 200 K	Up to 1 M
End-to-end latency (ms)	>50	<1
User mobility (km/h)	Up to 80	Up to 500
Peak data rate in cell (Gbit/s)	<1	>20
Traffic volume density (Tbit/s/km^2)	<1	Up to 10 s
User experienced data rate (Gbit/s)	<0.1	Up to 1

Table 1.
A comparison of the key parameters achieved in the 4G LTE mobile networks with similar parameters to be achieved in the 5G NR networks.

drawback for the implementation of the latter ones is the complexity and high cost associated with the need to lay the optical cables up to user terminals. In contrast, current wireless access networks that provide a flexible connection with a relatively simple infrastructure cannot meet growing in geometric progression demands to increase the capacity of mobile communication systems. The most promising technique to meet it, which is actively discussed in the referred publications, is to implement radio-over-fiber (RoF) network concept with FiWi architecture and to expend the operating frequency band up to millimeter waves (MMW) applying multi-position digital modulation on a radio-frequency (RF) carrier [25, 26].

Figure 3 shows a typical configuration of a RoF-based communication network including central station (CS) and set of remote (base) stations (RS), which are a key element of a RoF-based fiber-wireless fronthaul network (FWFN) that
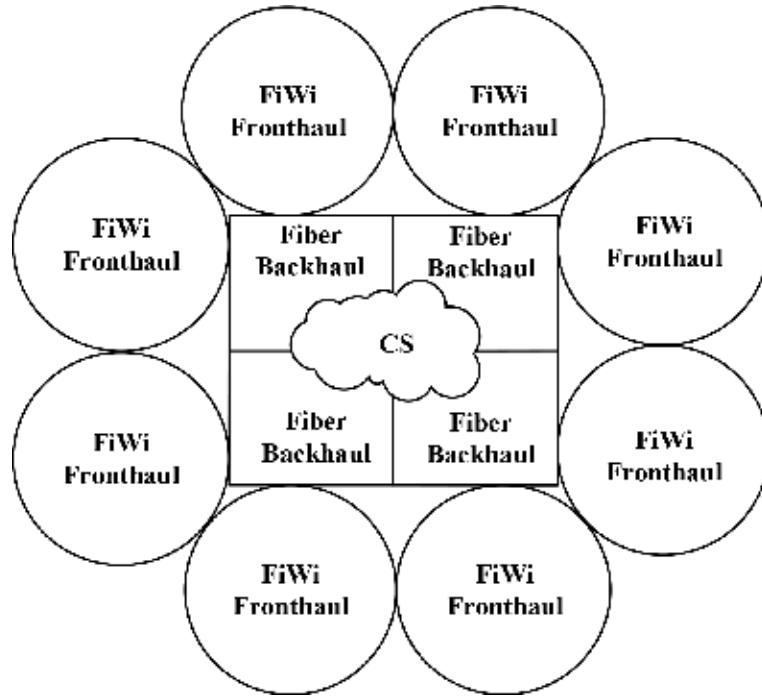
Figure 2.
Skeleton diagram for 5G NR cellular communication network.
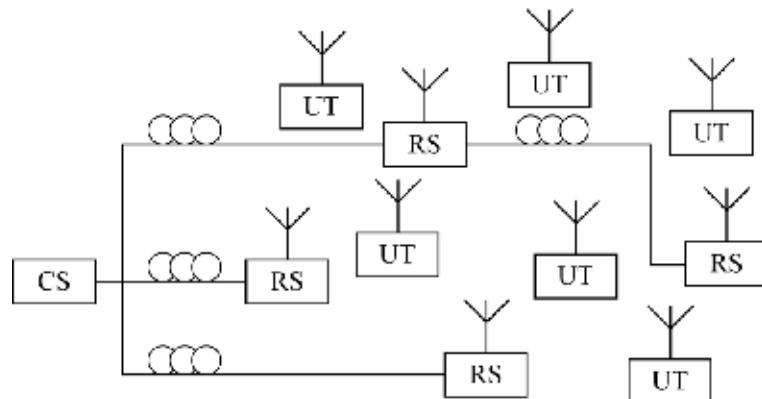
Figure 3.
A conceptual diagram of a RoF-based mobile communication network.

interactively (using downlink and uplink) connects the CS and each RS using fiber fronthaul links (FFL) and microwave or millimeter-wave band user radio terminals (UT). As is known, for the transmission of signals through a FFL, direct and inverse electro-optical conversions are required. The first one in the RoF-based communication system is usually performed with the help of an external electro-optical modulator (EOM) and the second with a photodetector.

In the framework of RoF concept, combining MMW band and FiWi network architecture inside FWFN is one of the promising ways to deliver intensive digital traffic with seamless convergence between wired optical backhaul and fiber-wireless fronthaul. In addition, FiWi technique allows converting directly a lightwave spectrum to MMW radio spectrum using a simple microwave-photonic-based up−/down-conversion scheme, which is important to keep the remote cells flexible, cost-effective, and power-efficient. **Figure 4** exemplifies a MMW-band FiWi architecture, in which CS is interactively connected with pico-cell's RSs through fiber-optic link. A typical position of RS is in the center of the service area; that is, for omnidirectional covering, four phased array antennas (PAAs) with an azimuth of 90° would be an optimal decision [16].

2.3 Example of a 5G NR network using small cell scenario and RoF concept

Let us illustrate small cell scenario using an example of building a backhaul network in a specific medium-scale city. **Figure 5** depicts 5G's backhaul fiber-optic network consisting of one macro-cell with service diameter of 5.3 km, inside of
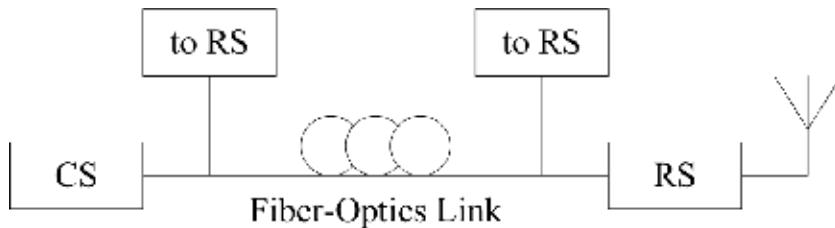
Figure 4.
A conceptual diagram of a RoF-based MMW fiber-wireless fronthaul network.
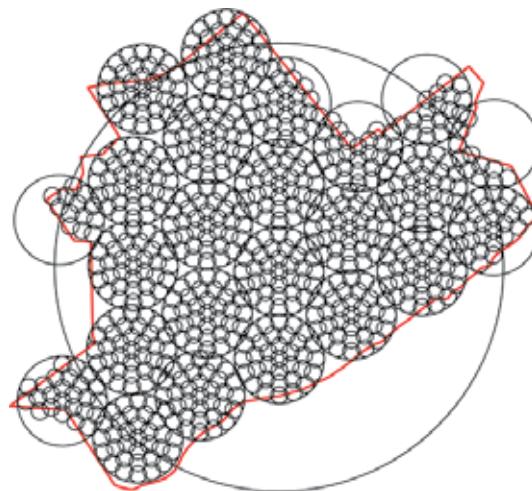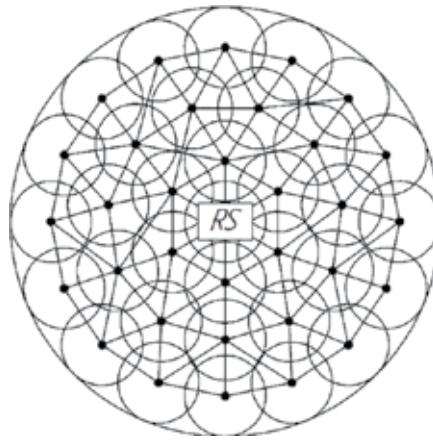
Figure 5.
5G's backhaul fiber-optic network of a medium-scale city.

Figure 6.
Conceptual diagram of pico-cells inside one micro-cell.

which 24 micro-cells of service diameter near 1 km are located. The red line marks the boundaries of the city.

Introducing a smaller partition, **Figure 6** shows a typical micro-cell diagram containing 33 pico-cells with a 200-m service diameter. In the center of each pico-cell, a remote station (RS) is located (see **Figure 3**), which is an interface between fiber and wireless network sections. A bold dot indicates it. All RSs are interactively interconnected via fiber-optic lines, forming a communication structure of the type "fully connected network."

2.4 Microwave-photonic-based approach

2.4.1 The area of microwave photonics

Microwave photonics (MWP) is a multidisciplinary research and industrial field encompassing optical, microwave and radio frequency (RF), and electrical researchers and engineers ([5, 20, 21, 26] and refs. cited there). This field in the last 30 years has attracted immense interest and generated many new R&Ds from both the scientific community and the commercial sector. Emerging applications for mobile communication network of FiWi architecture, sub-terahertz wireless systems, radar, and electronic warfare systems indicate that MWP is a subject of importance. By common opinion, MWP opens the way to superwide bandwidth characteristics at lower size, weight, and power as compared with traditional means [11]. For example, **Figure 7** depicts typical arrangements of MWP-based software-defined RF receiving (a) and transmitting (b) units. As it follows, a photonic circuit is inserted between two microwave electronic chains. For direct and inverse transformations of microwave and optical signals, there are two interfacing units at their bounds: electrical-to-optical (E/O) and optical-to-electrical (O/E) converters. Between the interfaces, there are various photonic processing units for switching, distribution, filtration, time delaying, and up/down frequency conversion of microwave signals in optical domain.

2.4.2 Millimeter-wave photonic technique in fiber-wireless-interfaced 5G wireless networks

To implement effective radio communication within small cell scenario, a number of leading countries developed a prospective spectrum including MMW bands up to 100 GHz (see **Figure 1**). As shown in numerous studies, MMW 5G network
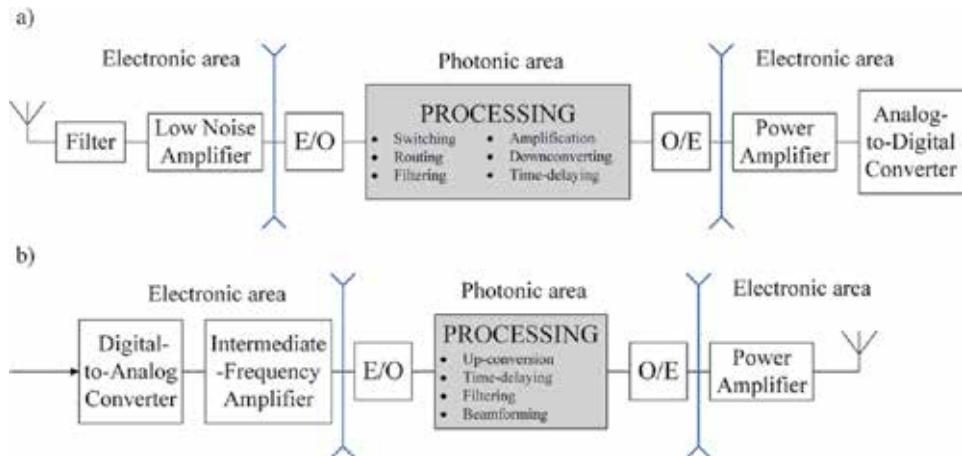
Figure 7.
A typical arrangement of MWP-based RF receiver (a) and transmitter (b).

infrastructure must be shared with a lot of small service zones controlled by the corresponding RS. In order to avoid inter-interference in these zones, one of the feasible approaches is to provide the RS with beam-steerable PAAs [16].

Generally, to form directional beams for transmission and receiving signals from adjacent UTs and RSs, MMW RS must use PAA with hundreds of antenna elements. In addition, FiWi technique allows converting directly a lightwave spectrum to MMW radio spectrum using a simple MWP-based up-conversion scheme [16], which is important to keep the remote cells flexible, cost-effective, and power-efficient and support seamless FWFN.

3. Design principles of fiber-wireless fronthaul network

In general, the fronthaul network of FiWi architecture represents the further development of cellular communication networks. The peculiarity of construction in comparison with the traditional system of cellular communication is in a much smaller area of cells down to pico-cells for mobile UTs with service diameter not more than 200 m and to femto-cells for indoor distribution with service diameter from tens of centimeters to 20–50 m. Due to the relatively small number of UTs inside the cell, it is critical to reduce cost of RS equipment, in fact, representing an effective interface between the optical and RF sections of the transmission system. The most promising solution to this problem is the ultimate simplification of the RS layout, which could be done by shifting all the processing procedures to the CS. If someone analyzes the diagram of **Figure 4** addressed to FWFN, from the functional viewpoint, two sub-systems are liberated that consist of optical distribution net-work (ODN) including CS hardware and fiber-optic link (FOL) and fiber-wireless interface (FWI) including a RS hardware and the same FOL. The proposed design principle is clearly illustrated in **Figure 8**.

The key advantages of the FiWi architecture for the communication networks are the following [1–4, 7]:

- Higher noise immunity, since data streams are mainly delivered through FOLs

- Small attenuation of signal power in fiber-based transmission path due to the fact that the losses in the fiber-optic cable are four orders of magnitude smaller than in the coaxial one
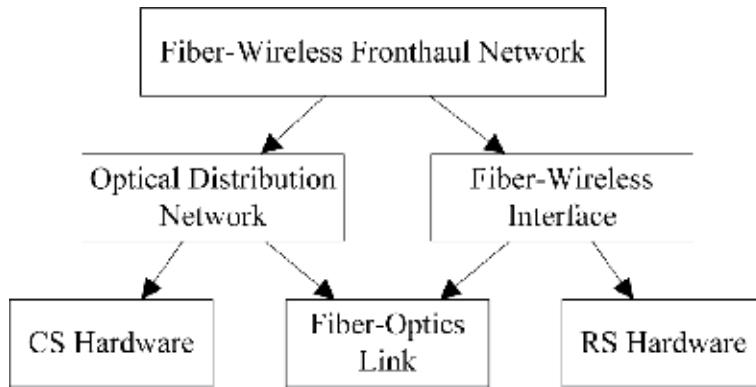
Figure 8.
The design principle of fiber-wireless fronthaul network.

- Relative simplicity of implementation and deployment at site by applying a remote base station concept that can service a significant number of wireless UTs

- A lower cost of construction and operation that simplifies the structure and reduces the power consumption of RSs due to using in the access networks the principle of transmission of digital streams on the carriers of the RF band

- Great future-proof design due to the fact that the ultra-wideband fiber-optic communication links guarantee minimal additional capital investments to upgrade the network throughput

Based on the benefits noted above, the next subsections review the principles, features, and ways to advance design of fifth-generation RoF-based access network using FiWi architecture.

3.1 Design of optical distribution network

As follows from the above discussion, the key function of the CS is the efficient electro-optical conversion. In this way, below we check by the simulation a CS for a RoF-based mobile network so to determine a feasible modulation method and the device for its realization when using optical transmission of multi-position quadrature amplitude modulation (QAM) of RF signals over FFL under investigation. Specifically, below we examine comparatively two key methods of optical modulation:

- Direct intensity modulation (DIM) for injection current of distributed feedback (DFB) laser or long-wavelength vertical cavity surface-emitting laser (LW-VCSEL). Hereinafter abbreviated as DIM-DFB or DIM-VCSEL, correspondingly.

- External intensity modulation (EIM) using electro-absorption modulator (EAM) or Mach-Zehnder modulator (MZM). Hereinafter abbreviated as EIM-EAM or EIM-MZM, correspondingly.

It should be noted that the key shortcomings of a DFB laser as compared to a LW-VCSEL are higher power consumption and substantially narrower band of

modulation [27]. On the contrary, the main preferences in terms of a signal transferring over an optical fiber consist in a smaller linewidth and a parasitic frequency modulation (chirp), which should lead to a significant extension in the permissible length of the FFL in the case of transmitting QAM signals. The quality is analyzed in terms of error vector magnitude (EVM) limit provided that the bottom of EVM value determined by the European Telecommunications Standard Institute (ETSI) corresponds to 8% for 64-QAM [28]. In all cases, the same 64-position QAM signal at the RF carrier in LR or HR (see **Figure 1**) will propagate over FFL.

We used the well-known commercial software VPIphotonics Design Suite™ as a tool for all computer simulation. Two key distortion sources are taken into account during the simulation procedure: a chirp of the lasers and modulators and a chromatic dispersion of the fiber. **Table 2** lists the common reference data for the FFL under study. In addition, **Tables 3** and **4** list the reference data for direct and external intensity modulation, correspondingly.

3.2 Proposed models and setups for simulation experiments

3.2.1 Direct intensity modulation

Figure 9 demonstrates VPIphotonics Design Suite's direct intensity-modulated FFL model and setup that contain the library models of DFB laser or LW-VCSEL, standard single-mode optical fiber (SMOF), and pin photodiode followed by the RF amplifier model. Their relevant parameters are in **Tables 2** and **3**. Besides, the setup includes the library model of DC source to control the DC bias current of a laser.

Parameter		Value
Length of pseudorandom bit sequence		$2^{15}-1$
Bit rate		2.5 Gbit/s
RF carrier frequency		1.8–10, 15, 40 GHz
Input RF power		−10 to −20 dBm
Type of RF modulation		64-QAM
Optical carrier		C-band (1552.52 nm)
Optical modulation		Intensity
PIN photodiode	Responsivity	0.9 A/W
	Dark current	100 nA
	3 dB bandwidth	50 GHz
	Optical input power	<3 mW
Post-amplifier	Gain	30 dB
	Noise spectral density	20×10^{-12} A/Hz$^{1/2}$
Optical fiber	Type	SMF-28e+
	Length	Up to 70 km
	Attenuation	0.2 dB/km
	Dispersion	$17e^{-6}$ s/m^2
	Dispersion slope	80 s/m^3

Table 2.
Common reference data for the FFL under study.

Parameter	DFB		LW-VCSEL	
	Value	**Reference**	**Value**	**Reference**
Operating current	60 mA	—	9 mA	—
Linewidth	300 kHz	[29]	4.5 MHz	[27]
Relative intensity noise	−150 dB/Hz	—	−160	[27]
Threshold current	8 mA	[29]	2.5 mA	[27]
Slope efficiency	0.15 W/A	—	0.23 W/A	[27]
Linewidth enhancement factor (α)	4.6	[30]	7.0	[31]
Adiabatic chirp factor (k)	3.2 GHz/mW	[30]	10 GHz/mW (at 1 GHz)	[31]

Table 3.
Reference data for direct intensity modulation.

Parameter	EAM		MZM	
	Value	**Reference**	**Value**	**Reference**
Operating voltage	−0.5 V	—	−5.7 V	—
Extinction ratio	14 dB	[32]	25 dB	[33]
Slope efficiency	0.14 W/V	—	—	—
Linewidth enhancement factor (α)	1.0	[34]	0 (X-cut)	[35]
Adiabatic chirp factor	0	[34]	—	—

Table 4.
Reference data for external intensity modulation.
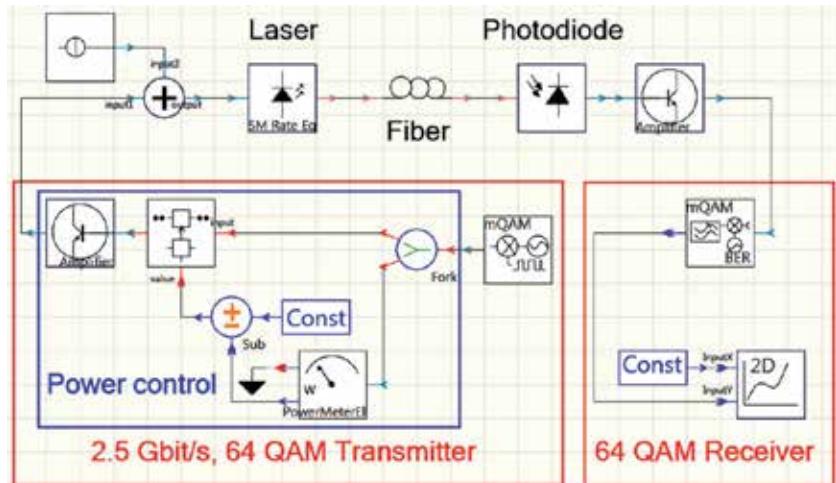
Figure 9.
VPIphotonics design suite's setup for a FFL with direct intensity modulation of QAM signals.

Finally, two instrumental library models are in the figure. The first one represents 2.5 Gbit/s, 64-QAM RF transmitter containing library model of QAM generator and output unit for power control. This module generates an electrical M-QAM signal up-converted at a desired frequency of the RF carrier. In addition, the second one represents electrical 64-QAM receiver. The module detects RF signal, decodes QAM signal, and evaluates the EVM of the QAM signal that has been transmitted. For

two-dimensional graphical representation of the data from the QAM receiver output, the model of numerical 2D analyzer is exploited.

3.2.2 External intensity modulation

Figure 10 demonstrates VPIphotonics Design Suite's external intensity-modulated FFL model and setup that contain the library models of EAM or MZM optically injected by C-band DFB laser. Their relevant parameters are in **Table 4**. Everything else in the figure coincides with the layout of **Figure 9**. Note that in the layout of **Figure 10**, the same DFB laser model as for direct intensity modulation is used, and its parameters are the same as in **Table 3** except the linewidth enhancement factor and the adiabatic chirp factor that are equal to zero.

3.3 Simulation results

Figure 11 depicts the examples of simulating EVM vs. fiber length characteristics for all devices under study when transmitting QAM-modulated 1.8-GHz RF carrier. The figure also illustrates constellation diagrams at EVM = 2% for all devices and at EVM = 5.7% for EIM-EAM. As it follows, MZM-based external modulation has the best values of EVM. Somewhat worse EVM characteristics are obtained by modulation using EIM-EAM and DIM-DFB, and the largest values of EVM are provided by the direct modulation using LW-VCSEL, which coincides with the known data [30, 36].

Studying the optical transmission of QAM signals at higher frequencies of the RF carrier using the FFL models of **Figures 9** and **10**, we found an interesting effect that was observed only in the case of direct modulation using a LW-VCSEL (**Figure 12**). This effect consists in decreasing the steepness of the distance characteristic of EVM with increasing RF carrier frequency, while the similar characteristics for the other devices under study remained the same as in **Figure 11**. The most probable reason for this atypical behavior is explained by the inverse frequency dependence of the magnitude and phase of the adiabatic chirp factor for LW-VCSEL [31].
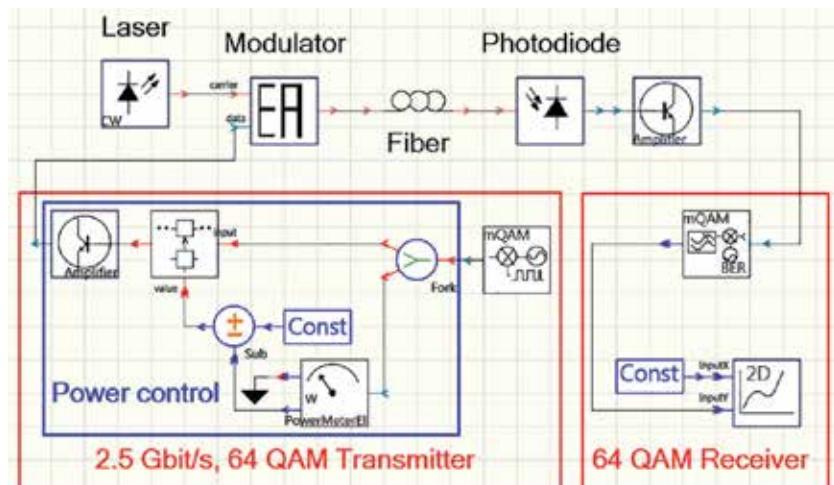
Figure 10.
VPIphotonics design suite's setup for a FFL with external intensity modulation of QAM signals.
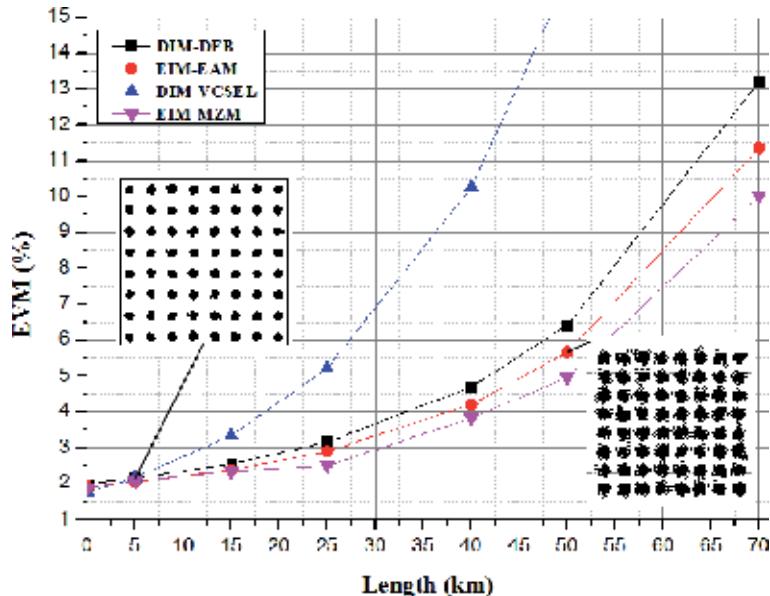
Figure 11.
Examples of simulating EVM vs. fiber length characteristics.
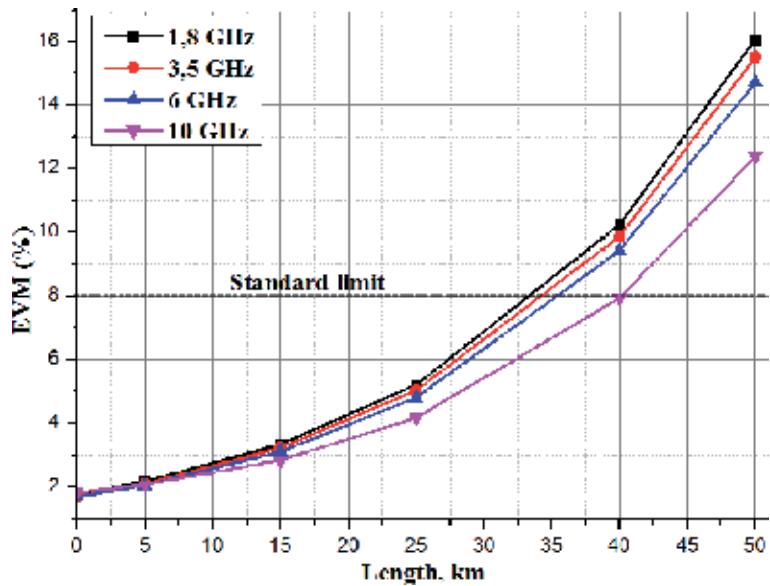
Figure 12.
DIM-VCSEL's EVM vs. fiber length characteristics at the various frequencies of RF carrier.

Some outcomes can be derived from the above analysis of the FFL under investigation:

- The allowable distance, when the standard EVM value during transmission of 64-QAM signal does not exceed 8% [28], is up to 62 km for an EIM-MZM, up to 58 km for an EIM-EAM, up to 55 km for a DIM-DFB, and up to 33 km for a DIM-VCSEL for the RF carrier of 1.8 GHz.

- At higher RF carrier frequencies up to 10 GHz, an atypical effect was detected for the direct modulation using a LW-VCSEL, which consists in a drop in the

slope of the EVM values with frequency increasing that is most likely explained by the presence of inverse frequency dependence of the adiabatic chirp factor's magnitude and phase.

Leveraging the study, below the results of the simulation experiment by the same computer tool imitating transmission of quadrature amplitude-modulated RF signals of 40 GHz (HR of **Figure 1**) or 15 GHz (IF band) through a FFL-connected CS and RS are discussed. Because the direct modulation bandwidth of modern laser sources does not exceed 10–15 GHz, this experiment is performed only for a circuit with external modulation using three types of EOMs: double-sideband MZM (DSB MZM), carrier-suppressed single-sideband MZM (CS-SSB MZM), and EAM. **Table 5** lists the reference data for the modulators under test.

The remaining data correspond to **Table 2** except for the number of modulation positions (16-QAM instead of 64-QAM) and data rates (1.25 Gbit/s instead of 2.5 Gbit/s), which are selected from the point of view of practical work in the MMW band. The VPIphotonics Design Suite's setup of the simulation experiment corresponds to **Figure 10**. **Figures 13, 14**, and **15** depict examples of simulating EVM vs. fiber length characteristics for the modulator under test during optical

Parameter	DSB MZM	CS-SSB MZM	EAM
Optical insertion loss	4 dB	6 dB	3 dB
Optical extinction ratio	20 dB	20 dB	14 dB
Slope efficiency	—	—	0.14 W/V
RF π-bias voltage	5.5 V	7.5 V	—
Electro-optical bandwidth	40 GHz	40 GHz	40 GHz
Linewidth enhancement factor (α)	0 (X-cut)	0 (X-cut)	1.0

Table 5.
Reference data for the modulators under test.
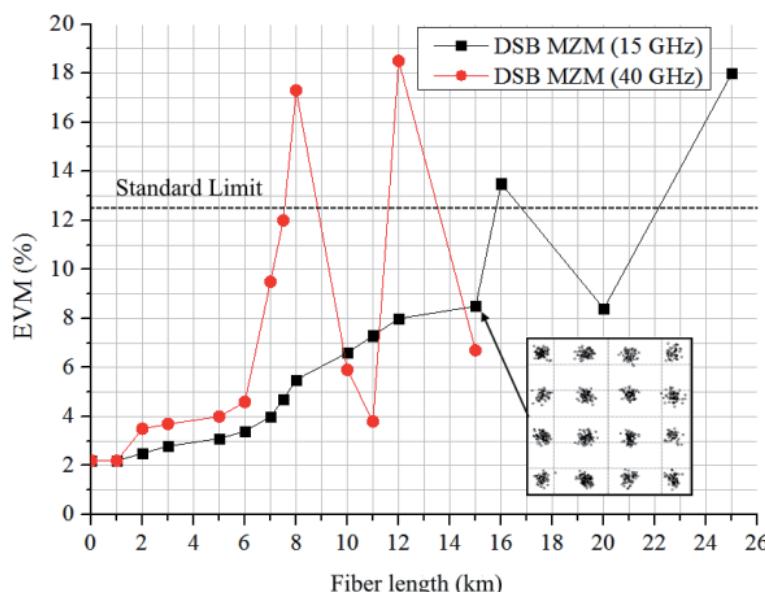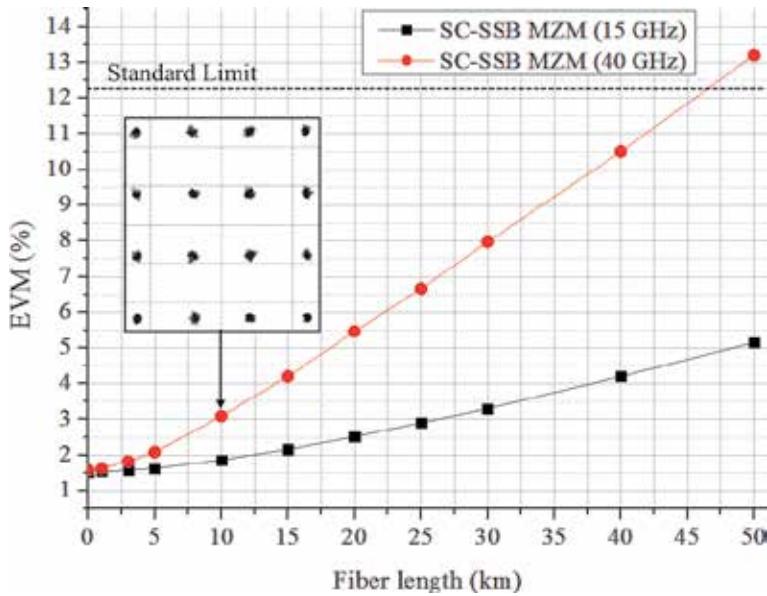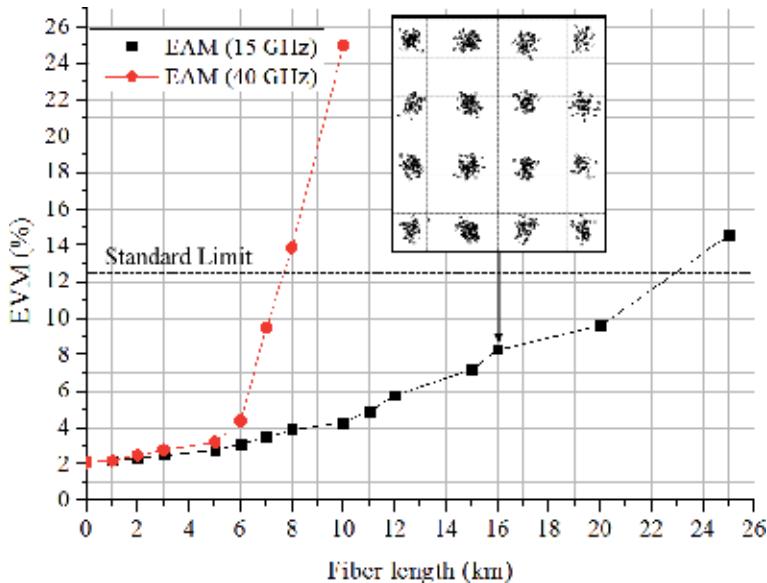
Figure 13.
Example of simulating EVM vs. fiber length characteristic for a FFL with DSB MZM under test.

Figure 14.
Example of simulating EVM vs. fiber length characteristic for a FFL with CS-SSB MZM under test.

Figure 15.
Example of simulating EVM vs. fiber length characteristic for a FFL with EAM under test.

modulation by 1.25 Gbit/s, 16-QAM, 15-GHz (transmission in IF band), or 40-GHz (transmission in MMW-band) RF signal, correspondingly. For the best vision, there are some insets in the figures showing constellation diagrams in specific points. In addition, in the figures, the dotted lines indicate the standard limit of the EVM during transmission of the 16-QAM signal, which is 12.5% [28].

The results of the simulation for the fiber-wireless fronthaul link of 5G NR system under study are summarized in **Table 6**.

The following outputs can be derived from our study:

- The minimum values of EVM were obtained for external modulation using CS-SSB MZM, which, nevertheless, requires the most complex control schematic, accordingly, and has the greatest value.

- The slope of the EVM characteristic increases with distance from SC-SSB MZM to DSB MZM through EAM, which corresponds to the known data [30, 34].

- The significant fluctuations in the values of the EVM at the 40-GHz RF carrier (see **Figure 13**) are characterized by the effect of dispersion in an extended optical fiber. To eliminate it in order to increase the length of the FFL, it is required to introduce at its end a dispersion corrector, which is a standard element in a fiber-optic communication system.

3.4 Design of fiber-to-MMW-band wireless interface

An important element of a RoF-based mobile communication network (see **Figure 3**) is a remote station, through which an interactive fiber-wireless interface is implemented. Recently, we have proposed and previously investigated advanced design concept of cost- and power-efficient base station for emerging FiWi networks, in which for a multifrequency conversion of a RF carrier, a MWP-assisted optical frequency comb generator (OFCG) based on an optical recirculation loop (ORL) technique using two SC-SSB optical modulators was developed [19]. Leveraging the application of this OFCG for a realistic case, the simulation results by the same computer tool imitating multiwavelength optical frequency comb generation and transmission of quadrature amplitude-modulated RF signals through OFCG-based FWI of a FiWi-architected RS are discussed.

Figure 16 shows the VPI model and setup for simulation of the OFCG scheme under study. There are four units depicted in the figure: the composed model of ORL includes library models of optical X-coupler, SC-SSB modulator, optical amplifier (OA), and optical band-pass filter (OBPF), library models of continuous-wave semiconductor laser (CW-SL) emitting at the frequency ν_0 as an optical source, RF generator (RFG) as a RF signal source, and library instrumental model of optical spectrum analyzer (OSA). In order to close the ORL, the output of OBPF through the service unit T and input of SC-SSB are connected to X-coupler's port "input2" and port "output2," correspondingly. During the simulation, RFG acts as a source of the reference RF signal (f_{ref}), while using the OSA, the output optical spectrum is recorded.

Figure 17 shows the VPI model and setup for simulation of OFCG-based fiber-to-MMW-band wireless interface, while transmission of QAM-modulated RF signals is supported. The scheme represents the downlink channel of FiWi-architected

Device under test	RF carrier (GHz)	Allowable distance of FFL (km)
DSB MZM	15	16
	40	7.3
SC-SSB MZM	15	Much more than 50
	40	47
EAM	15	23
	40	8

Table 6.
The results of the simulation.
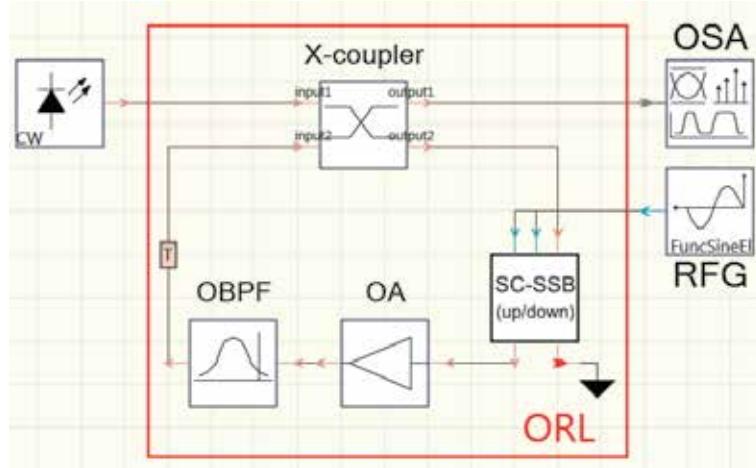
Figure 16.
VPI model and setup for simulation of the OFCG.
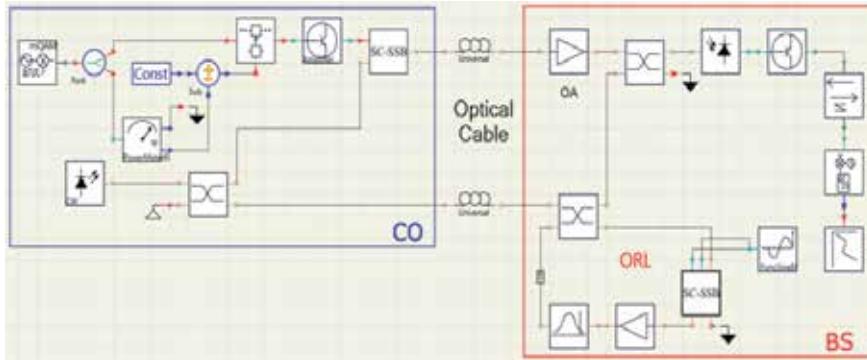
Figure 17.
VPI model and setup for simulation of OFCG-based fiber-to-MMW-band wireless interface.

RoF system and consists of three units imitating the operation of CS, RS, and two-fiber optical cable between them. The CS includes the same laser model, the radiation of which is divided into two branches using a Y-coupler, library model of SC-SSB modulator with suppressing lower sideband and library instrumental model of QAM RF transmitter. The latter contains library models of QAM generator and output unit for power control followed by electrical amplifier. This module generates an electrical M-QAM signal up-converted at a given RF carrier frequency. The optical cable includes two equivalent library models of single-mode optical fiber. Such a remote optical feed reduces the cost of the RS. Besides the OFCG model (see **Figure 16**), the RS includes library models of optical amplifier, X-coupler, photodiode, and electrical post-amplifier outputted to the model of QAM RF receiver (see Section 3.1).

In the course of the research, first of all, the possibility of creating a multifrequency OFCG with the closest arrangement of the teeth is checked. Then, the transmission quality of a digital RF signal with multi-position QAM through the downlink channel of the RS using fiber-to-MMW-band wireless interface is analyzed. **Table 7** lists the common reference data for the OFCG under study. The reference data for the fiber-to-MMW-band wireless interface under study correspond to **Table 2**, with the exception of the frequency of RF carrier (37–43.5 GHz instead of 1.8–40 GHz) and data rates (1.25 Gbit/s instead of 2.5 Gbit/s).

Figure 18 demonstrates an OSA's spectrum of multiwavelength optical frequency comb output following the setup of **Figure 16**. As one can see from the figure, the OFCG under study includes 21 optical carriers with the spacing of 0.3 GHz and a level nonuniformity of less than 5 dB.

The results of the simulations are presented in **Figure 19**. For a clear view, there are some insets in **Figure 19** showing constellation diagrams in specific points. In particular, as one can see from the figure, due to dispersion in the optical cable, the EVM values increase with a slope of near 0.17%/km reaching a standard limit for 64-QAM of 8% [28] at a distance of 40 km.

The following output can be derived from our study. When transmitting digital radio signals with 64-QAM on millimeter-wave RF carriers (37–43.5 GHz), even when using SC-SSB MZM and high coherent laser, the dispersion in an optical cable has a significant impact on the quality of the received signal. However, the error is within the standard limit up to a distance of 40 km.

3.5 Studying an optimal signal transmitting RoF-based fiber-to-MMW-band wireless interface

Finally, we consider and discuss the optimal design principle of an interactive fiber-wireless fronthaul network when distributing digital radio signals over fiber-optic link. The feasible variants are compared in **Table 8**. For the possibility of quantitative analysis, we take the widely used bitrate for the modern networks of 1 Gbit/s (e.g., the standard Gigabit Ethernet).

Parameter	Value
Laser source frequency (ν_0)	193.3 THz
Laser linewidth	10 kHz
Reference RF frequency (f_{ref})	0.3 GHz
Type of modulator inside optical recirculating loop	SC-SSB (up/down)
Gain of recirculating loop (g)	0.8 < g < 1
Number of up or down round trips	Not less than 10
Level nonuniformity of output comb teeth	Not more than 5 dB

Table 7.
Reference data for the OFCG under study.
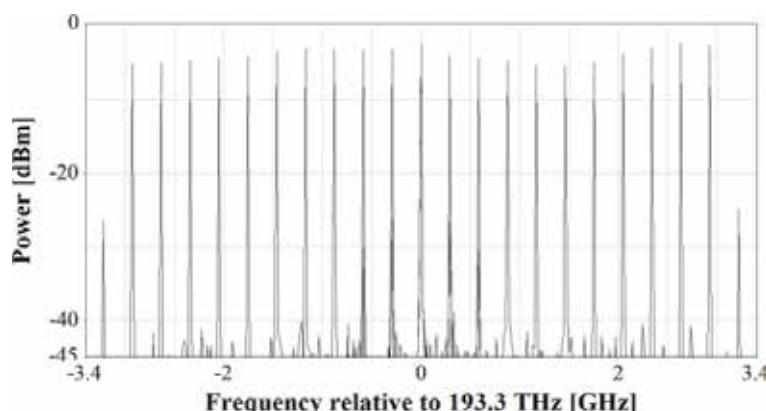
Figure 18.
A spectrum of multiwavelength optical frequency comb output.

Based on the general comparison, below the results of our recent investigations to design optimally a fiber-wireless fronthaul including a central station, MMW wireless interface, and pico-cell remote station are presented.

In particular, we compare the three options of distributing signals through fifth-generation fronthaul communication network of fiber-wireless architecture with a wireless section operating in MMW band: in baseband, in intermediate frequency band, and directly in RF band on the same MMW frequencies. **Figure 20** demonstrates three possible options including interactive transmission in the baseband (a), in the IF band (b), and in the RF (MMW) band (c). The following abbreviations are used in the figure: TSL, tunable semiconductor laser; EOM, electro-optic modulator; PD, photodetector; RFM, RF modulator; RFDM, RF demodulator; MIMO, multiple input/multiple output; IFM, IF modulator; RFC, RF converter.

To verify the optimal layout, a transmission quality simulation of a 64-QAM, 2.5 Gbit/s digital signal transmitted at a frequency in the IF band (15 GHz) or in the
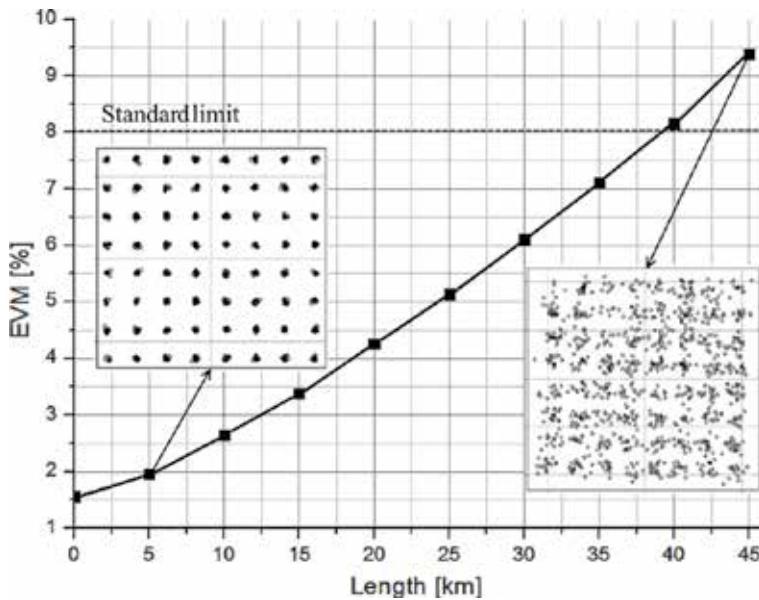
Figure 19.
EVM vs. optical cable length.

Transmission range	Option 1. In baseband	Option 2. In the band of intermediate RF signals	Option 3. In the band of RF carriers
Type of FOCS	Digital	Analog	Analog
Upper modulation frequency (GHz)	1	10–15	40–80
Relative bandwidth (%)	100	40	30
Demands to signal-to-noise ratio	Low	High	High
Demands to the equipment linearity	Low	High	Middle
Complexity of CO layout (cost)	Low	Middle	High
Complexity of BS layout (cost)	High	Middle	Low

Table 8.
Comparison of the feasible options for transporting signals over FWI.
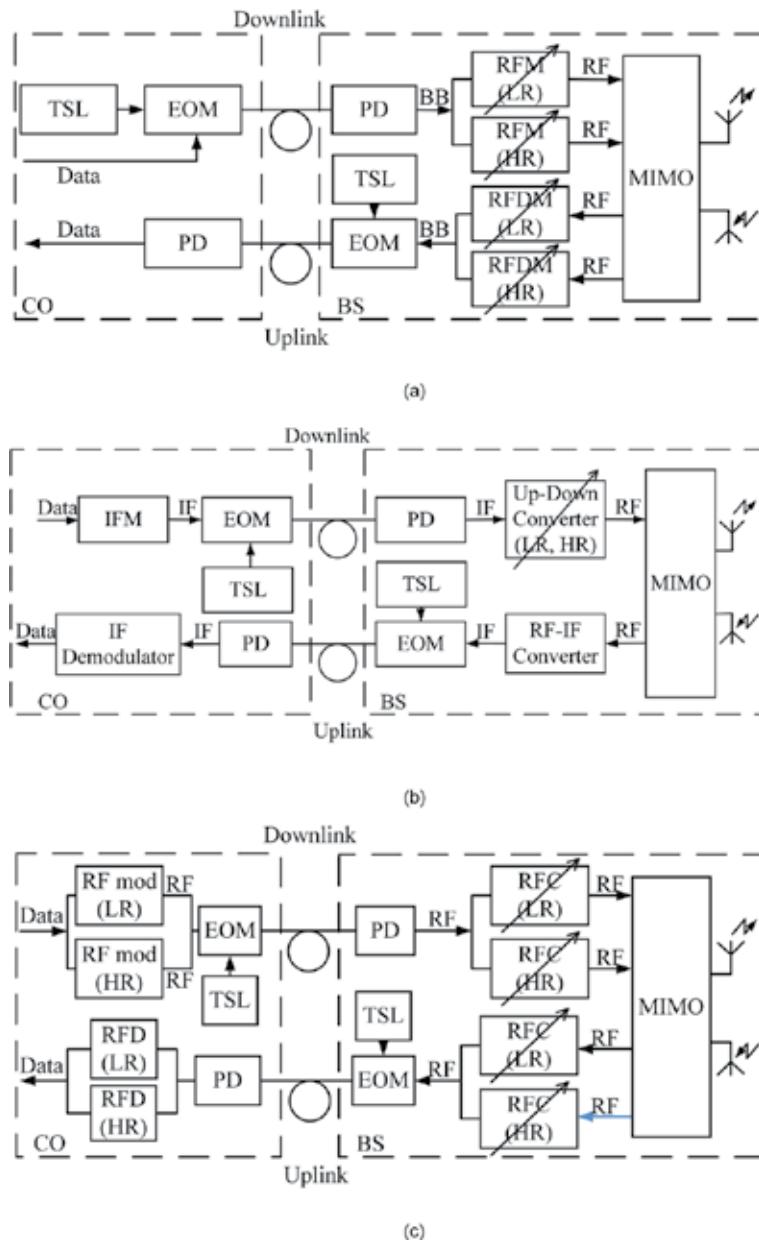
Figure 20.
The possible options of transmitting signals through FiWi fronthaul. (a) Baseband-over-Fiber transmission, (b) IF-over-Fiber transmission, and (c) RF-over-Fiber transmission.

MMW band (40 GHz) through FFL using CS-SSB MZM, was performed by the same off-the-shelf computer tool of VPIphotonics Design Suite. The reference data used in the calculations are taken from **Table 5** for the modulator and from **Table 2** for the entire FWFN. The result, which is a dependence of the EVM vs. the fiber length, is shown in **Figure 21**.

As follows from the figure, the transmission at 40 GHz is carried out at a much worse quality than at 15 GHz. In particular, the standard for 64-QAM limit of 8% [28] is achieved in the first case with a fiber link length of 23 km and as much as 60 km in the second case.
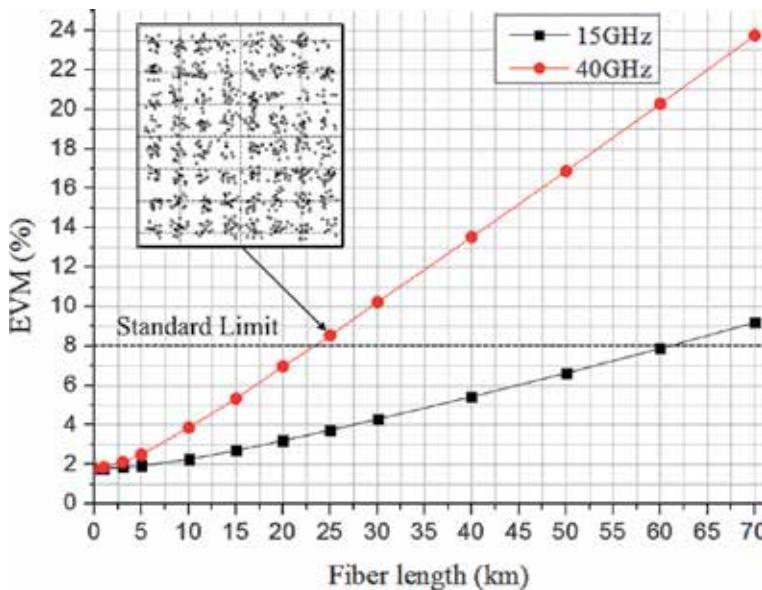
Figure 21.
EVM vs. fiber length characteristic.

The following outcomes can be drawn:

- To realize Option 1, it is necessary to use digital fiber-optic communication system, whereas for the second and third options, analog system is required with its inherent higher requirements for the signal-to-noise ratio and the linearity of the equipment.

- The value of the upper modulation frequency in the second and, especially, in the third options is significantly higher in comparison with the first one, which tightens the requirements for the electronic and optoelectronic components of the CS equipment and, as a result, its cost.

- The relative bandwidth of the transmission channel for the second and third options is substantially lower than in the first one, which simplifies the circuitry of the CS and RS equipment's amplifying and converting units and, as a result, improves their cost characteristics.

- The option with transmission in the RF carrier band is realized with the least number of transformations on the RS, which minimizes its cost and, consequently, the cost of the entire user access network. However, the fiber-optic transmission of the MMW-band signals has a serious limitation due to the dispersion effect of standard optical fiber.

We believe that the optimal approach would be IF-over-fiber transmission in spite of this option requiring an interface at the RS that has to perform RF up- or down-conversion. Nevertheless, transmission in IF band (see **Figure 20(b)**) provides versatility, as there is a simple possibility of RF conversion at a RS, both in the LR and in the HR (see **Figure 1**). For the effective implementation of it, we have proposed and described two RS schemes [14, 19] capable of frequency converting both LR and HR.

4. Conclusion

In the chapter, we proposed and highlighted in detail the specific for incoming fifth-generation mobile communication system principles to optimally design an access network using small cell scenario, radio-over-fiber concept, and microwave-photonic-based approach. In the generally accepted interpretation, the small cell scenario means a consistent network partitioning into micro- and pico-cells with a service area diameter of not more than 200 m. Radio-over-fiber concept is to design pico-cell network based on fiber-wireless architecture. Using microwave-photonic-based approach means the formation and processing of transmitted radio signals in the optical range, which leads to a significant improvement of the bandwidth features at lower size, weight, and power as compared with traditional characteristics of network equipment. The main idea behind the proposed principle to design pico-cell fiber-wireless fronthaul network is to split it into two sub-systems that consist of optical distribution network including central station hardware and fiber-optic link and fiber-wireless interface including a remote station hardware and the same fiber-optic link. In order to verify efficiency of the proposed design principles, we performed modeling in a well-known computer-aided design environment VPIphotonics Design Suite. The goal of the study was to examine and select optimal modulating scheme and transmitter parameters to propagate higher-order quadrature amplitude modulation signals at radio-frequency carriers of millimeter-wave band over radio-over-fiber-based fiber-wireless Fronthaul network using advanced commercial optoelectronic devices and standard single-mode optical fiber. In the result of simulation experiments, optimal design principles of optical distribution network, fiber-wireless interface, and fiber-wireless fronthaul network as a whole have been proposed, described, and validated. Particularly, the study of the optimal method and device for transmitting multi-positional QAM signals at RF carriers showed that in the so-called low range (see **Figure 1**), the maximum allowable distance of a fiber-optic link is provided up to 33 km for a LW-VCSEL and up to 55 km for a DFB laser in the case of direct modulation, as well as up to 58 km for an EAM and up to 62 km for a SC-SSB MZM in the case of external modulation. In addition, in the so-called high range (see **Figure 1**), the maximum allowable distance of a fiber-optic link is significantly reduced, reaching at best not more than 23 km even when using a SC-SSB MZM, which, nevertheless, requires the most complex control schematic, accordingly, and has the greatest value.

Acknowledgements

Conflict of interest

The authors declare the lack of the "conflict of interest."

Author details

Mikhail E. Belkin*, Tatiana N. Bakhvalova and Alexander S. Sigov
Scientific and Technological Center "Integrated Microwave Photonics",
MIREA-Russian Technological University, Moscow, Russian Federation

*Address all correspondence to: belkin@mirea.ru

IntechOpen

References

[1] Andrews JG, Buzzi S, Choi W, Hanly SV, Lozano A, Soong ACK, et al. What will 5G Be? IEEE Journal on Selected Areas in Communications. 2014;**32**(6):1065-1082

[2] Chen S, Zhao J. The requirements, challenges and technologies for 5G of terrestrial mobile telecommunication. IEEE Communications Magazine. 2014;**52**(5):36-43

[3] Munn J. Our 5G future: In the fast lane with numerical simulation. Microwaves & RF. 2016;**16**:48-50

[4] Frenzel L. Making 5G happen. Microwaves & RF. 2017:1-5

[5] Browne J. What role will millimeter waves play in 5G wireless systems? Microwaves & RF. 2018;**10**:38-42

[6] Waterhouse R, Novak D. Realizing 5G. IEEE Microwave Magazine. 2015;**16**(8):84-92

[7] Boccardi F, Heath RW, Lozano A, Marzetta TL, Popovski P. Five disruptive technology directions for 5G. IEEE Communications Magazine. 2014;**52**:74-80

[8] Novak D, Waterhouse R. Emerging disruptive wireless technologies—Prospects and challenges for integration with optical networks. In: IEEE: 2013 Optical Communication Conference and Exposition and the National Fiber Optic Engineers Conference (OFC/NFOEC). Anaheim, CA, USA. 17–21 March. 2013. pp. 1-3. DOI: 10.1364/OFC.2013.OTu3E.2

[9] World Radiocommunication Conference 2019 (WRC-19). Available at: https://www.itu.int/net/events/eventdetails.asp?eventid=14719 [Accessed: 23 September 2019]

[10] ITU. Final acts of WRC-2015 World Radiocommunication Conference.

Resolution 238, Geneva 2016. pp. 296-298

[11] Novak D et al. Radio-over-fiber technologies for emerging wireless systems. IEEE Journal of Quantum Electronics. 2016;**52**(1):1-11

[12] Al-Raweshidy H, Komaki S, editors. Radio over Fiber Technologies for Mobile Communications Networks. Norwood: Artech House; 2002. 436 pp

[13] Sauer M, Kobyakov A, George J. Radio over fiber for picocellular network architectures. IEEE Journal of Lightwave Technology. 2007;**25**(11):3301-3320

[14] Belkin ME. The building principles of a cost- and power-efficient base station for emerging fiber-wireless networks. In: International Conference on Microwaves, Communications, Antennas and Electronic Systems, COMCAS 2017; 13–15 Nov. 2017; Tel Aviv, Israel. pp. 1-4

[15] Belkin ME, Golovin V, Tyschuk Y, Sigov AS. Studying an optimal approach to design combined fiber-wireless telecom systems. In: 41st International Conference on Telecommunications and Signal Processing, TSP 2018, Athens, July 4–6. 2018. pp. 43-46

[16] Belkin ME, Fofanov D, Golovin V, Tyschuk Y, Sigov AS. Design and optimization of photonics-based beamforming networks for ultra-wide mmWave-band antenna arrays. In: Array Pattern Optimization. United Kingdom, London: IntechOpen; 2018. pp. 47-67 Available from: https://www.intechopen.com/online-first/design-and-optimization-of-photonics-based-beamforming-networks-for-ultra-wide-mmwave-band-antenna-a

[17] Fofanov DA, Bakhvalova TN, Alyoshin AV, Belkin ME, Sigov AS.

Studying microwave-photonic frequency up-conversion for telecom and measurement equipment. In: IEEE Radio and Antenna Days of the Indian Ocean (RADIO2018), Mauritius. 2 pp.

[18] Bakhvalova T, Belkin M, Fofanov D. Advances in fiber-wireless network architecture approach to the next-generation communication systems. In: Proceedings of the Seventh International Conference on Advances in Computing, Communication and Information Technology—CCIT; 27–28 October 2018. Rome, Italy. 2018. pp. 62-67

[19] Belkin ME, Bakhvalova T, Turitsyn S, Sigov A. The design principles of reconfigurable versatile base station for upcoming communication networks. In: 26th Telecommunications Forum (TELFOR2018), Belgrade, Serbia. Nov. 2018. pp. 180-182

[20] Fofanov DA, Bakhvalova TN, Alyoshin AV, Belkin ME. Microwave-photonics frequency up-converter for telecom and measurement equipment. IOP Conf. Series: Materials Science and Engineering. 2019;**524**:-012006

[21] Belkin M, Bakhvalova T. Microwave photonics and fiber-wireless interface— A principal approaches to design millimeter-wave network equipment for the next-generation wireless communication systems. In: Proceedings of the eighth International Conference on Advances in Computing, Communication and Information Technology—CCIT, 23–24 April 2019, Birmingham, UK. pp. 27–31

[22] Bakhvalova TN, Fofanov DA, Alyoshin AV, Belkin ME. Fiber distribution networks with direct and external modulation by digital QAM-signals. In: Proceedings of the 42nd International Conference on Telecommunications and Signal Processing (TSP2019), 1–3 July 2019. Budapest, Hungary; 2019. pp. 241-244

[23] Belkin M, Bakhvalova T. Studying optical frequency comb-based fiber to millimeter-band wireless interface. In: The Fifteenth Advanced International Conference on Telecommunications (AICT-2019), 28–31 July 2019. Nice, France; 2019. pp. 94-98

[24] Belkin M, Bakhvalova T, Sigov A. Studying an optimal approach to distribute signals through fiber-wireless Fronthaul network. In: IEEE Conference on Microwaves, Communications, Antennas and Electronic Systems (COMCAS-2019), November 4-6, Tel Aviv, Israel. 2019. in press

[25] Dat PT, Kanno A, Kawanishi T. Radio-on-radio-over-fiber: Efficient fronthauling for small cells and moving cells. IEEE Wireless Communications. 2015;**22**:67-75

[26] Dat PT, Kanno A, Umezawa T, Yamamoto N, Kawanishi T. Millimeter- and terahertz-wave radio-over-fiber for 5G and beyond. IEEE Photonics Society Summer Topical Meeting. August 2017:21

[27] Belkin ME, Belkin L, Loparev A, Sigov AS, Iakovlev V. Long wavelength VCSELs and VCSEL-based processing of microwave signals. In: Pyshkin S, Ballato J, editors. Optoelectronics— Advanced Materials and Devices. Croatia: InTechOpen; 2015. pp. 231-250 Chapter 6

[28] ETSI. Minimum requirements for Error Vector Magnitude. In Technical Specification, LTE; Evolved Universal Terrestrial Radio Access (E-UTRA); User Equipment (UE) radio transmission and reception (3GPP TS 36.101 version 14.3.0 Release 14), ETSI, 2017-04, p. 215

[29] Laserscom LDI-1550-DFB-2.5G Data sheet. Available at: www.laserscom.com

[30] Kim BG, Bae SH, Kim H, Chung YC. RoF-based mobile Fronthaul networks implemented by using DML and EML

for 5G wireless communication systems. IEEE Journal of Lightwave Technology. 2018;**36**(14):2874-2881

[31] Halbritter H, Shau R, Riemenschneider F, Kogel B, Ortsiefer M, Rosskopf J, et al. Chirp and linewidth enhancement factor of 1.55 μm VCSEL with buried tunnel junction. Electronic Letters. 2004; **40**(20):1266-1268

[32] Optilab EML. OL5158M-LC Data sheet. Available at: www.optilab.com

[33] iXblue Photonics LiNbO3 modulator MXER-LN-10. Available at: https://ph otonics.ixblue.com/

[34] Salvatore RA, Sahara RT, Bock MA, Libenzon I. Electroabsorption modulated laser for long transmission spans. IEEE Journal on Quantum Electronics. 2002;**38**(5):464-476

[35] Koyama F, Iga K. Frequency chirping in external modulators. IEEE Journal of Lightwave Technology. 1988; **6**(1):87-93

[36] Mikroulis S, Karabetsos S, Pikasis E, Nassiopoulos A. Performance evaluation of a radio over fiber (RoF) system subject to the transmitter's limitations for application in broadband networks. IEEE Transactions on Consumer Electronics. 2008;**54**(2):437-443

Section 4

Electronic Circuit Design

Chapter 9

Design and Analysis of Analog to Digital Converter System Clock Source Using Direct Digital Synthesizer

Desmond Tung and Rosmiwati Mohd-Mokhtar

Abstract

A requirement of multiple format standards by mobile telecommunication (GSM, CDMA, WCDMA, and TD-SCDMA) test set needs to be delivered possibly at lower cost. As to support its capability, phase-locked loop (PLL) frequency synthesizer has been designated as an essential part in most of the design within the box. The old design may be bulky and subject to many issues with the components' variation and aging effect. In recent years, the direct digital synthesizer (DDS) has been popularly in used to replace the PLL architecture. This chapter will focus on the DDS selection, architecture topology, prototyping, implementation technique with both hardware and software, and performance as a clock source to a sampled system as referred to receiver interest. The key parameters in the sampled system greatly rely on the jitter and phase noise specification. If they are not properly defined, the overall signal-to-noise ratio (SNR) at the sampled system output will be impacted. Eventually the receiver quality will be degraded and resulted in tremendous loss. Thus, a proper reconstruction filter design will be delivered to ensure the jitter and phase noise performance is met without degrading the existing specification by taking accountability into the matching characteristic and signal integrity.

Keywords: direct digital synthesizer, clock source, analog to digital converter, jitter, filter design

1. Introduction

Many mature products in the market utilize PLL frequency synthesizer to provide a referencing element either in radio frequency or digital application. During the beginning, there were not many options or topologies available in designing one. As a result, the released product is often found to be bulky and expensive considering the tight tolerance and accuracy requirement. As time flies by, many researchers start to explore into a possibility to implement the frequency synthesizer digitally, and there the direct digital synthesizer (DDS) was introduced [1–3].

Figure 1 shows the basic DDS block diagram. The DDS in modern days could come with multiple capabilities [4]. All of them were integrated together to provide a complete solution to customer with reasonable price. For instance, a complete DDS has both DDS core and digital-to-analog converter (DAC) integrated into
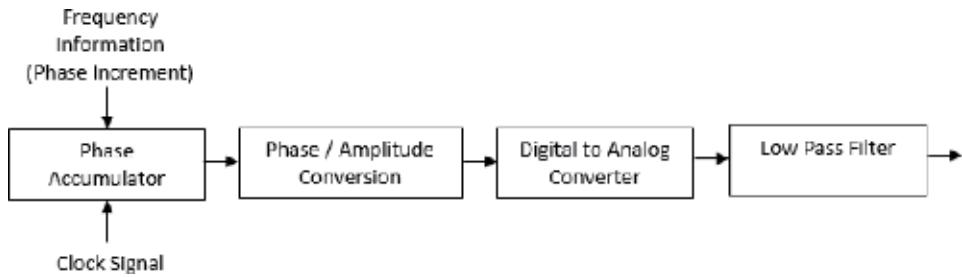
Figure 1.
General DDS block diagram.

single package. Of course, these will not come true without the advances in IC fabrication technology. In addition, a process architecture like system in package (SiP), system on a chip (SoC), and 3D stacked die also helps in providing DDS as a solution to complex design such as modulators, local oscillator (LO) clock, and chirp generator. Company like Analog Devices who has been the leader in DDS market offers a dedicated DDS IC with minimum requirement of digital signal processing (DSP) that fit the customer's need, while Xilinx, Altera, and other companies offer the solution through programmable synthesis. The trade-off of price and space requirement in a design must be made.

By programming the DDS, the desired frequency hopping, numerous modulation formats, and data rates can be achieved. Even digital modulation is possible since the processing signal is in digital. Hence, the adaptivity and flexibility of DDS make it ideal for not only radio frequency but also for many applications. Since it is digital, the thermal drift susceptibility is the least concern to a design which is favorable to most of the designers. Basically, DDS can be found in a lot of applications such as cellular base stations, wireless local loop base stations, cellular phone, and test and measurement equipment [5].

The DDS is a mixed signal device that can be analyzed as digital and analog. The DDS itself is digital, while DAC and reconstruction filter are considered as analog. In FPGA implementation, the whole design can be coded and synthesized. The DDS often comes in small package that perfectly fits in replacing bulky circuitries. The cost is significantly lower than a conventional PLL design. The only drawback is that an additional reconstruction filter is required to shape the discretized output into a comprehensive sinusoidal reference. The study will focus on the available topologies in designing a replacement clock generator with specified specifications.

As to ease the design, some of the methods were researched as explained in [6]. A tool like ADISIMDDS can be used as preliminary justification but not the actual implementation. This method suggests a proper conduct of frequency planning in order to get a usable output spectrum to fit the application. It helps to ensure the performance of DDS-based radar application.

There are many hybrid designs in wireless communication system that used DDS to work with PLL. The reason is due to that the DDS has a high-frequency resolution, has accuracy, and is easy to be implemented in PLL process control [7]. Besides, the remarkable fast frequency switching within the guard time enables time-division multiple access (TDMA) dynamic channel assignment possible [8]. The only problem with DDS is really the output spectrum. It requires a filter to pair with it, ensuring the spectrum is cleaned due to the harmonics generated upon frequency generation.

A complete DDS in the market comes with phase adjustment capability. It allows the design to associate self-adjustment knowingly to the phase delay inherited by the filter. At high frequency, the phase response may be worst due to the PCB

production process [9]. The result is a measure of jitter and phase noise that degrade the system performance [10]. These parameters are important to high-speed digital link as well as ADC system [11]. Considering a system uses DDS as global clock distribution, a time-varying noise (jitter) will impact the time available for logic operation between sequential elements [12] causing the system to fall apart.

There are some challenges that are anticipated during the research and development cycle. The constraints like tight specification in a one to one replacement, suitable techniques and topologies in clock source implementation, and phase noise and jitter specification needed to be straighten out during the concept and investigation checkpoint. Next, upon the development checkpoint, interfacing between DDS and reconstruction filter might be subjected to uncertainty. The mismatch element among components and interfacing circuitry may be a problem as well. Thus, a proper conduct of design and simulation needs to be performed.

2. Implementation and methods

The prototype was done with the aid of MATLAB, Simulink, and Altera Quartus II software. Nevertheless, Altera DE2 development kit was the hardware used. The advantage of having prototyping with FPGA is because the output can be observed by adjusting the internal module such as phase accumulator bits, look-up table design, and optimization to identify the minimum acceptable performance. From there, the Analog Devices DDS selection can take place. Once the DDS has been selected, the Analog Devices evaluation board (AD9852) will be used to evaluate the DDS. At the same time, the filter design and simulation can be performed.

2.1 Prototyping with Altera FPGA and MATLAB Simulink

Resource occupation of two look-up tables (LUTs) or read-only memory (ROM) would not meet the FPGA available resources. Thus, the approach of ROM reduction is necessary. By downsizing the LUT and shrinking the data content in the ROM, the required LUT has now reduced, down to one as shown in **Figure 2**. One solution is to replace one of the LUTs with K-bit inverters. The idea is to invert the output from the LUT as the phase accumulator reaches half of its counts. In Simulink, there is no direct inverters that can be used; however, with the addition of XOR gate, it is possible.

The whole idea of downsizing is to adjust the content stored inside the ROM. By reducing the maximum value to $\pi/2$, the steps can be remained, and therefore, the frequency resolution remained. These will change the ROM from the half-wave data
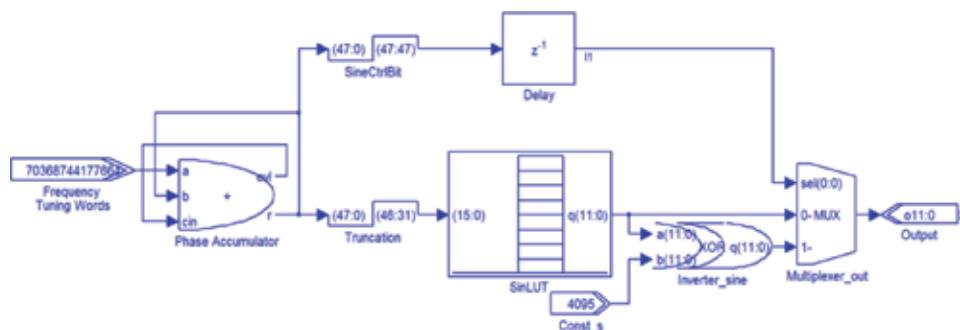
Figure 2.
Simulink model 17 bit DDS with single LUT half wave.

into quarter-wave data with successfully reducing the address bit of ROM by 1 bit, from 16 bits down to 15 bits.

2.2 Reconstruction filter design and simulation

A DDS takes two inputs, a reference frequency (f_{REF}), and tuning word M (refer to **Figure 3**). The built-in multiplier allows DDS to output a wider range of frequencies from a low reference frequency. Tuning word is essential to the DDS as it specifies the jump step of fine-resolution output frequency. The shorter the tuning word, the lower the frequency.

The reconstruction filter is needed to filter off the unwanted spurs that come with the synthesized output. By using a low reference frequency that is available in the current ADC sampled system, the multiplier within the DDS will be used. The product of multiplication can be added onto the output spectrum. Secondly, the output from LUT or ROM is driving the DAC. DAC output spectrum is anticipating the primary phase truncation spur and the second and third harmonics. **Figure 3** shows the DDS with reconstruction low-pass filter (LPF).

The specification has been defined as shown in **Table 1**. Agilent ADS simulation tool was used to design the reconstruction filter. With the spurious spectrum induced by DDS design itself, the steepness of the transition band is important to the overall quality of the performance. Elliptical response is preferable as it has a sharpen roll-off with lower order than the Butterworth and Bessel although both
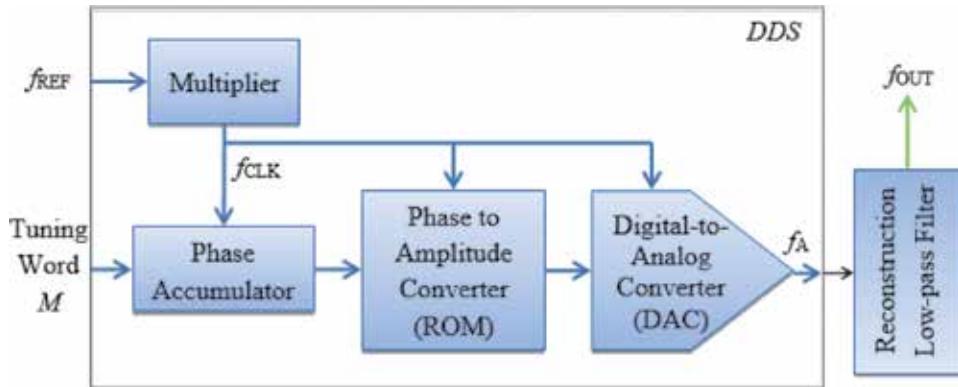
Figure 3.
DDS with reconstruction LPF.

Criteria	Target	Buffered
Cutoff frequency (f_{pass})	40 MHz	42 MHz
Passband ripple (R_{pass})	0.01 dB	0.01 dB
Stopband gain (A_{stop})	70 dB	78 dB
Response	Elliptical	Elliptical
Input/output impedance	50 Ω	50 Ω
Number of order	7	7

Table 1.
Targeted reconstruction filter specification.

offer maximal flat passband. In addition, since the comparator will be used to convert the sine wave-filtered output into square wave, an occurrence of low amount of ripple in passband will not impact the output performance.

As to utilize the internal high-speed comparator that comes with AD9852, two LPFs are required. The stopband gain for the filter specification is originally 70 dB with a cutoff frequency of 40 MHz. To cater for the tolerances and anticipated standardization of the passive component value loss between the two LPFs, a buffered margin of 8 dB (>10%) of stopband gain and 5% of cutoff frequency that is 2 MHz are added to the specification. Basically the buffered margin of 8 dB is essential due to de-normalized value to nominal value conversion loss. The result may end up with 70 dB stopband gain that is still within the specification.

On the other hand, the criticalness of the tolerance contributed by the components needs to be addressed. The total effect of conversion loss and tolerances may cause the filter to operate out of specification. The original design targeted a 40 MHz cutoff frequency. By simulating the tolerances using ADS Monte Carlo analysis, it is found out that 40 MHz gave a tight margin that even the temperature coefficient has not been included. The preliminary prototype specification is based on 78 dB stopband and 40 MHz. A total of 100 iterations is programmed in ADS Monte Carlo analysis to solve and estimate the worst-case scenario of the filter response when the tolerances are included.

In LPF design, basically the number of order is obtained with the help of MATLAB *fdatool*. With the aided design, a seventh-order LPF is advised in designing a LPF that meets the specification. Once the minimum order has been defined, designing the passive LPF can be conducted as shown in **Figure 4**.

2.3 Using AD9852A complete DDS as clock source

There are basically few steps needed to operate AD9852A as a clock source for 8960 ADC. Before that, the hardware configuration of DDS must take place. The simplified configuration is shown in **Figure 5**. In the design, there are two distinct power supplies being used to power the chip. They are digital VDD and analog VDD. The reason of having two power supplies for operation is to achieve a better isolation. This will minimize the risk of signal integrity concern such as simultaneous switching noise (SSN) or PLL ringing due to complex design within the system. As for both analog and digital ground, they are tied together.

The plan in this work is to make use of internal reference multiplier due to low reference differential clock on ADC (20 MHz) that connects to REFCLK and
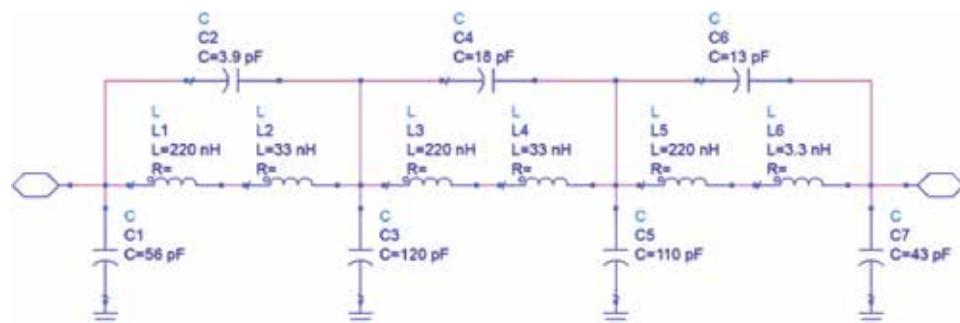
Figure 4.
Seventh-order elliptic LPF.
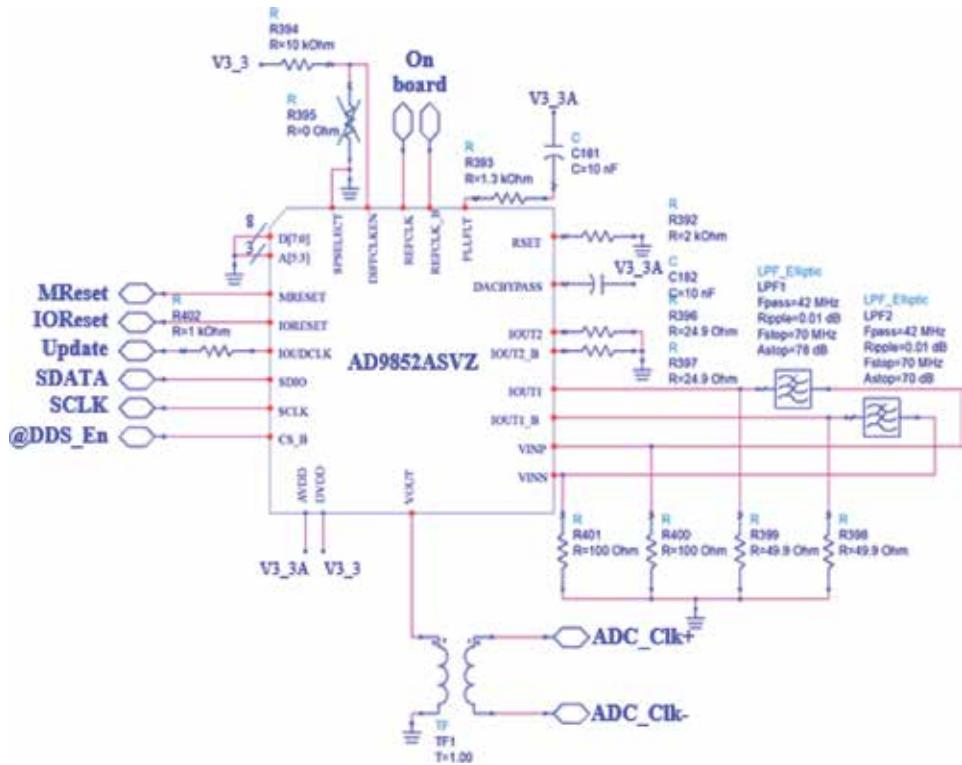
Figure 5.
Simplified DDS implementation schematic diagram.

REFCLK_B. Therefore, the internal PLL will be used. There is this pin PLL FILTER (PLLFLT) that serves as zero compensation that requires to be connected as close as possible to AVDD through RC network of 1.3 kΩ and 10 nF [13]. Nevertheless, the differential clock enable pin (DIFFCLKEN) was pulled up for differential clock application. There is also a pull-down resistor in place with no load condition. This allows future changes if the reference clock is to be used as single-ended instead of differential.

In order to improve the harmonic distortion and spurious-free dynamic range (SFDR), a bypass capacitor to AVDD is connected to the DDS's DACBYPASS pin. This pin can be left unconnected, but it may degrade the output performance slightly. Besides, there are two output channels that come with AD9852 composed of a cosine DAC and a control DAC. However, only one channel is required. Although the control DAC can be disabled programmatically, it is necessary to terminate them making sure it will not interfere with adjacent pins. With that, 24.9 Ω resistors are used for termination. The output strength of the current DACs can be controlled by the design through an external resistor.

Both channel one outputs are connected through two LPFs that have been designed in the previous section. By having both outputs filtered, a clean zero-crossing to the comparator will yield a better output. Before feeding the filtered output into the comparator, there is another 100 Ω resistor termination. They are there to ensure a 50% output duty cycle from the comparator not affected by differential input setup.

To save the pin count, serial programming has been designated in implementation. Connecting SPSELECT pin to the ground tells AD9852 to operate with serial

programming. For that, eight data bits and three address bits from MSB have connected to the ground to prevent AD9852 misbehave. With these configurations, only six control lines are needed to operate the AD9852 DDS. Basically, upon master reset, the IO update clock pin (IOUDCLK) by default is configured as an output. To avoid bus contention or any damage to the DDS, a 1 kΩ resistor is connected in a series. Even if the update pin from the host controller changes state, a constant current of 3 mA is drawn. Ultimately the output VOUT from the DDS will be connected to ADC through a balun that converts the balanced clock into unbalanced that is composed of differential output [14].

2.4 Characterization and qualification test profile

Once the first article of DDS is being produced, the characterization will take place to ensure that the design meets the specification. It can be benchmarked against the previous design to understand if the desire characteristic is achieved or otherwise. As soon as it is ready, hardware and software qualification will be started. It is meant to qualify the hardware to come out with new set of specification that could be better or tighter than the previous design. Similarly to software qualification, parameter like stability, bugs, and speed should be taken care upon, interfacing and controlling the hardware.

By replacing the DMP PLL with DDS, the characterization needs to be done thoroughly. As previously mentioned, apparently the phase noise and jitter are the major concern for an ADC sampled system. Tuning speed on the other hand will enhance user experience by speeding up the measurement. The spurious-free dynamic range is another measure of the clock signal quality. Also, since the circuit operates around other components, even with proper fence and shielding, it would be good to understand the cross-talk effects.

Agilent 8960 has a guaranteed specification that is operable at 55°C. Hence, the temperature test profile must be planned carefully. The temperature profile was used to characterize and qualify the newly developed DDS. The same temperature profile will be used upon characterization prior to specification roll-up as well as qualification. Ambient temperature within the factory is kept at 23°C. Hence, a ramping and soaking period of 30 minutes is required to allow that the temperature settled within an error band of ±1°C. Respective test point will be connected through a semirigid coaxial 50 Ω cable along with a 1 meter cable connection between the DDS output and the PXA or DSA. A set of workbench test script will be used to collect the data with previous test line limit. The actual test line limit will only reflect the DDS design after qualification. **Table 2** is constructed as the required sampling clock in Agilent 8960 ADC board.

Basically the qualification cycle will be performed using Weiss WK 1600 chamber series. As shown in **Figure 6**, there will be a switchbox being deployed to switch against different equipment. The reason being is that the probe of each instrument has its own impedances. Without the switchbox, all the measurement will become invalid as the impedances may be altered. After all, the master controller, a host will be doing the entire necessary configuration to control the equipment and perform data collection.

Agilent PXA with phase noise option feature was used to characterize the phase noise performance of DDS. The center frequency was adjusted accordingly to the frequency of interest. Phase noise is usually expressed in decibels relative to carrier power per unit Hz bandwidth (dBc/Hz). An array list of offset frequency was setup by decade to measure the phase noise. Besides of phase noise measurement, the SFDR can be measured to anticipate the strongest spur power level after the fundamental frequency.

Formats	Measurement	Sampling rate (MHz)
GSM GPRS EGPRS	PFER/PvT/EDPOW/EvM/EPOW	18.9583
	GSM IQ tune	19.9899
	EPSK IQ cal	13
	EPSK XTXP	19
	EPSK SPC	20
TD-SCDMA	TD ACLR/SEM	19.2
	TD POW/CLPC/DPOW/FDT/OLPC/OOS/TOOP/WQ/WQ cal	15.36
	TD IF cal/OBW	12.8
WCDMA	WOBW	20.48
	W-AVP	20
	W-SEM	23.04
CDMA	C-audio	16
	C-POLT	10
	C1x-other measure	14.7456
IS136	Min	2.5
	Max	40

Table 2.
Wireless mobile communication sampling frequency specification in 8960.
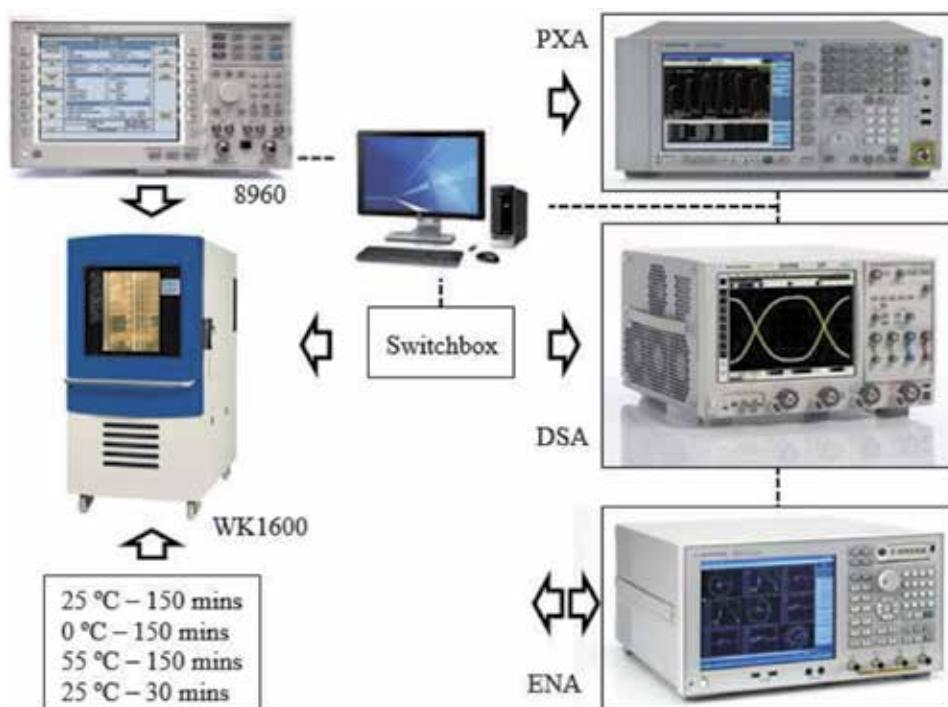
Figure 6.
Test and measurement setup.

3. Result and analysis

3.1 Software simulation

Table 3 summarizes the outcome from the prototypes. Apparently the ROM downsizing has effectively reduced the required memory bits by 80% (1–483,840/ 2,396,160). However, notice that by reducing the half wave from two LUTs into one may reduce the required memory but not the price. If Cyclone II is to be used, a wider user I/Os of 422 that has the same memory as of Cyclone III with 196 user I/ Os is recommended but costly.

Cyclone III may seem reasonable, but notice that the available user I/Os is far lesser than Cyclone II. At around the same price, device EP2C35F484C6 provides more user I/Os than EP3C40F324C6. With successful ROM downsizing, device EP2C35F484C6 can be used, and the output performance remains as of first proto-type. The advantages in terms of cost, user I/Os, and output performances have provided a satisfied outcome.

The LPF designs' comparison is shown in **Table 4**. By using two simulation tools along with their available feature, the LPF has been successfully designed. The specification is guaranteed as close as the original specification. The responses provided by MATLAB are giving the designer an idea of how the filter supposed to behave. The hand-calculated value was validated by ADS design guide. After fitting with nominal value, a worst-case analysis was done using Monte Carlo analysis by accounting all the component tolerances within the filter itself. Comparatively the final design has a promising specification at passband around 40 MHz. The only disadvantage is that the stopband gain is closed to the initial unbuffered specifica-tion. However, as expected, nominal value conversion will definitely introduce these losses, and hence, trade-off has been made.

Design	Cyclone	Simulink recommended device (User I/Os)	Minimum memory bits required	Used	Usage	Price per unit
Half wave 2 LUTs	III	EP3C55F484C6 (328)	2,396,160	1,572,864	66%	$ 212.30
Half wave 1 LUT	III	EP3C40F324C6 (196)	1,161,216	786,432	68%	$ 122.20
Half wave 1 LUT	II	EP2C70F672C6 (422)	1,152,000	786,432	68%	$ 387.20
Quarter wave 1 LUT	II	EP2C35F484C6 (322)	483,840	393,216	81%	$ 149.60

Table 3.
Comparison among different prototypes.

Criteria	Target	MATLAB	ADS design guide	Final
3 dB cutoff ($f_{cut\text{-}off}$)	42 MHz	46.2 MHz	46 MHz	46.06 MHz
Gain at 40 MHz	0 dB	−0.006 dB	−0.004 dB	−0.003 dB
Passband ripple (R_{pass})	0.01 dB	0.01 dB	0.01 dB	0.01 dB
Stopband gain (A_{stop})	−78 dB	−78 dB	−78.07 dB	−71.561 dB

Table 4.
LPF designs comparison.

3.1.1 DDS clock generator performance

Figure 7 displays the overall DDS simulation model as clock generator. By setting the DAC output current to be greater at about 20 mA, the output performance has a better slew, jitter, and phase noise performance. **Figure 8** shows example of the DDS performance at frequency of 2.5 MHz. From simulation, the higher the frequency, the LPF phase shift became obvious. As the frequency keeps increasing, the effectiveness of the LPF is noticeable as the high-frequency components from the source are being filtered off. With all the changes, the output differential clocks are still centered at 1.6 V at 1.5 V_{P-P}. Signals are the square wave by assuming the DDS output is exactly half of system clock.

As frequency increases, all nonlinearity effects have been revealed. From 2.5 to 40 MHz, obvious distortion is observed at the output from DDS. The same occurred to the filtered output. At the frequency of 14.7456 MHz, output lags its input by almost 90°. As frequency goes to 19.9899 MHz, the phase lags increased beyond 90°. When it comes to 40 MHz, the phase lag increased to approximately 270°.

The phase lag analysis shows that there is some discrepancy between the s-parameter response results against transient analysis. This is possible as the marker that is placed on transient analysis is bounded by the number of points. To increase the point resolution, longer time is required to perform the simulation. However, based on the percentage, the validation of less than 4.5% is achieved, and therefore, the LPF phase response has been verified.
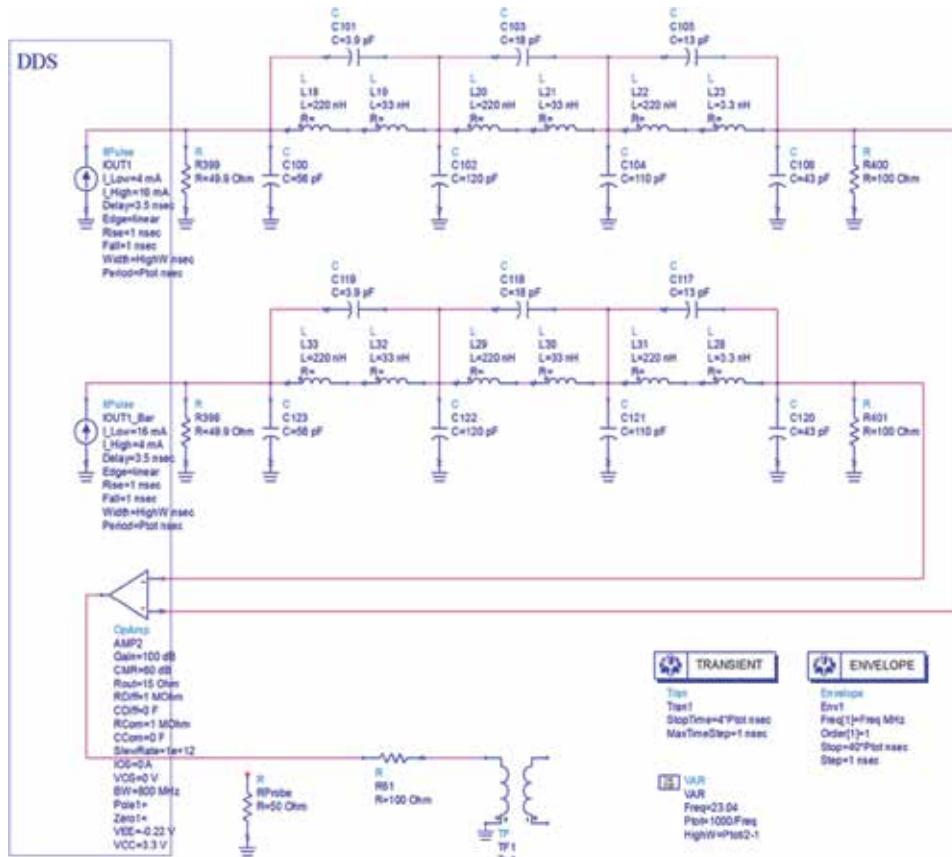
Figure 7.
DDS simulation model as clock generator.
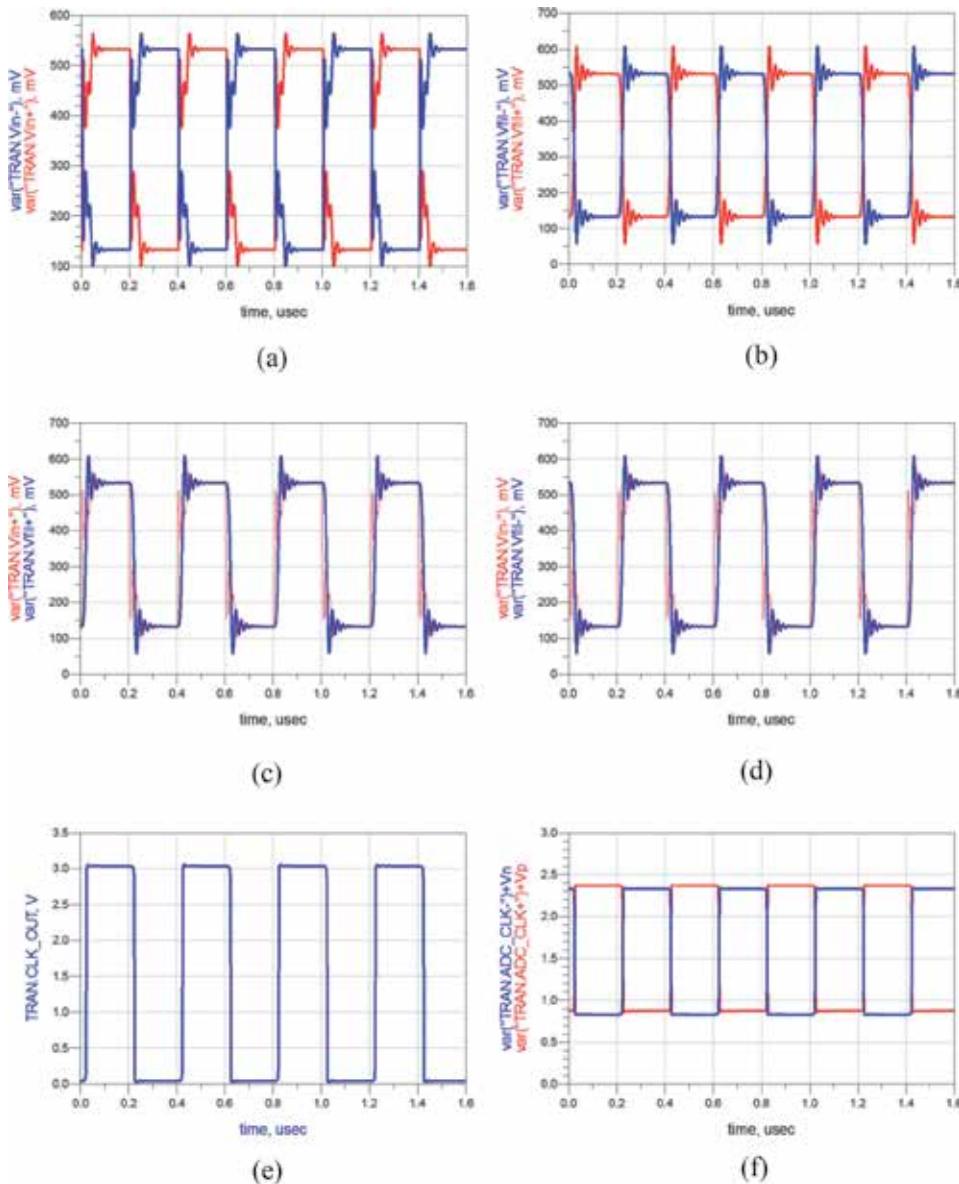
Figure 8.
DDS performance at 2.5 MHz (IS136). (a) DDS output: IOUT1 (red), IOUT1' (blue), (b) Filtered output: IOUT1 (red), IOUT1' (blue), (c) Phase response comparison: Vin+ (red), Vfil+ (blue), (d) Phase response comparison: Vin- (red), Vfil- (blue), (e) DDS clock output, and (f) ADC differential clock.

3.2 Hardware implementation

In this work, DE2 is being used as prototype implementation. Choosing it is essential as it comes with all the required peripherals and it is also inexpensive. The design was enhanced by downsizing the ROM, making DE2 a perfect fit. Apparently by having DE2, working alone is impossible to collect the time domain measurement. Therefore, an external 12-bit DAC AD7545A was used to perform the analog conversion. The basic setup of the DAC is by having the VREF connects to a 0.5 V so that the output will rail from 0 to 1 V. Together with feedback circuitry and a 50 Ω output, VOUT1 can be measured using scope.

During a 2.5 MHz output (**Figure 9(a)**), a smooth sine wave is observed. This is because the PA is extremely slow and outputting a relatively fine resolution at the
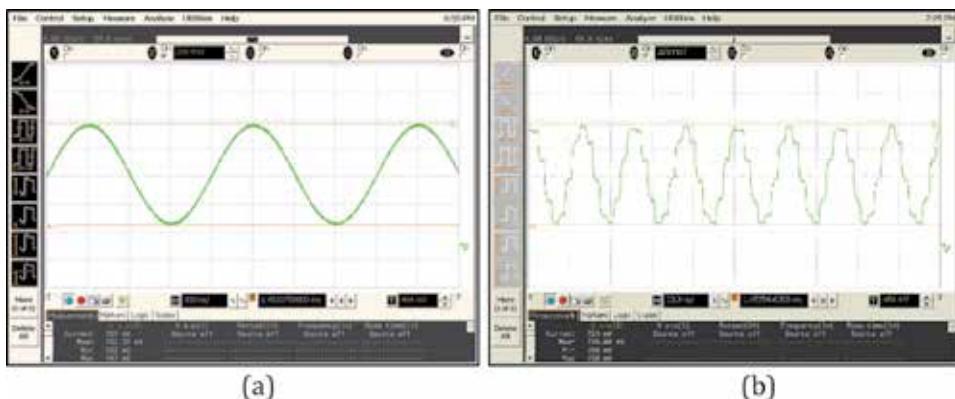
Figure 9.
DE2 quarter wave DAC output at (a) 2.5 MHz and (b) 40 MHz.

output. However, as the frequency changed from 2.5 to 40 MHz (**Figure 9(b)**), the staircase shape output starts to reveal. Hence, the comparison between simulation and practical measurement can be conducted. The results yield the same thing as of the simulated outcome. The same output response has been measured at AD9852A DDS output.

The LPF was characterized throughout the temperature profile running from 0 to 55°C. This is meant to qualify the LPF in terms of temperature coefficient introduced by passive components. **Figure 10** demonstrates a collective data for all samples. It helps to understand part-to-part variation together with temperature impact towards the LPF. From **Figure 10**, it seems that the designed filters were rather stable and consistent over the passband. The 3 dB gain can be found around the 42 MHz which follows the design specification. There is a lot of fluctuation that can be seen beyond the stopband frequency of 70 MHz by design. These variations will only contribute towards the phase response at the output of LPFs. Since the output magnitude is greatly attenuated, the phase changes will not impact DDS performance at all. By accounting part tolerances, temperature coefficient, and PCB properties, the stopband gain greater than 70 dB has been achieved. From the
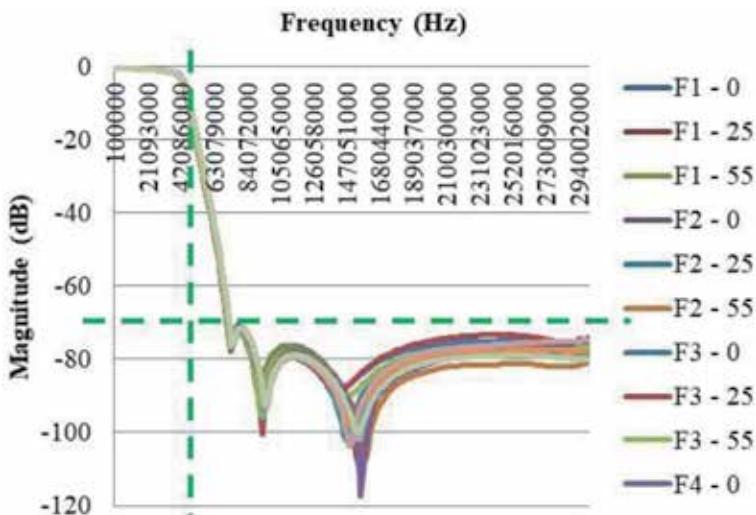
Figure 10.
Comparison among five samples across temperature.

results, the design of LPFs is meeting the specification and guaranteed stable with wide range of operating temperature.

3.2.1 DDS clock generator performance

There are five measurements being conducted for DDS clock source characterization. They are spectrum, phase noise, spurious-free dynamic range, jitter, and transient measurement. **Figure 11** shows the frequency domain measurement result at 14.7456 MHz, and **Table 5** tabulates the measured power level across temperature. It seems that over the operating frequency range, there is about 1.54 dB loss. This loss can be found at lowest frequency of 2.5 MHz. The reason for this loss may be due to the bandwidth of the comparator itself. Ideally during simulation, the comparator is expected to operate from DC up to 400 MHz that is greater than the system clock of 300 MHz. However, in practical, the bandwidth could start somewhere from 1 to 300 MHz, and therefore, a lower power level has been obtained around the corner. Beside this lowest frequency, the rest of them behave as expected. Additional 2 dB shall be attained when the load impedance is increased to a maximum of 1 MΩ that is exactly the impedance of next stage digital circuits.

The phase noise is quite consistent over temperature (**Figure 12**). From the result, it indicates that at sampling frequency of 2.5 and 20 MHz, the phase noise is as stable as little changes across temperature. On the other hand, starting from 10 to 19.9899 MHz, there is about a 3 dB shift from 0–55°C, at frequency offset of 1 kHz. The same goes to 23.04 MHz as well as 40 MHz.

The spurious-free dynamic range has only being qualified at the ambient temperature. The results that were discussed are chosen from those that are closest to the −60 dB baseline (example of 16 MHz frequency is shown in **Figure 13**). The strongest spur level has been found to be −60 dB below or 70 dBc within a frequency range from 13 to 20.48 MHz. This satisfied the wide-operating range to ensure the integrity of DDS clock source. This third harmonic spur is expected and managed to suppress the elliptic LPF since it fell out of the 42 MHz range.
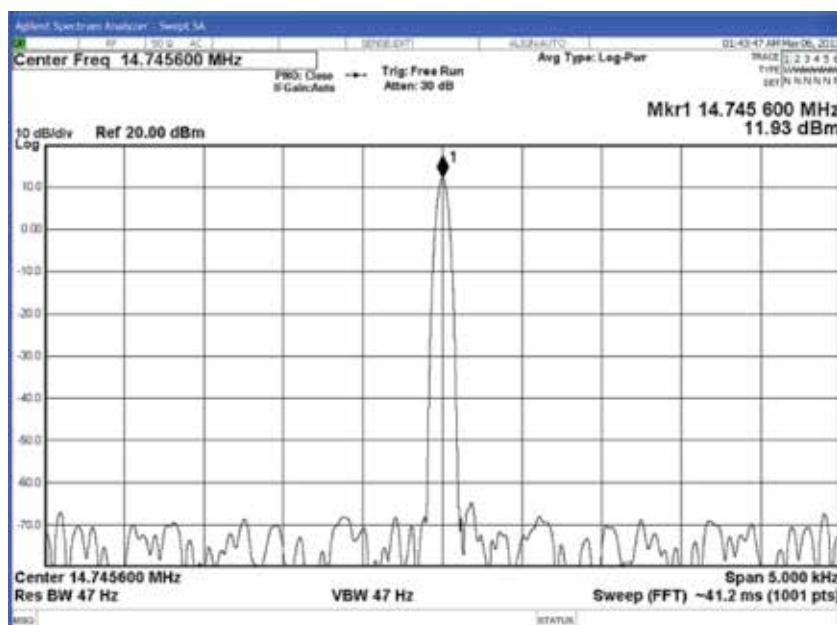
Figure 11.
Frequency domain measurement result at 14.7456 MHz.

Sampling rate (MHz)	Measured (0°C)	Measured (25°C)	Measured (55°C)
2.5	10.75	10.39	10.39
10	11.98	11.95	11.93
12.8	11.98	11.95	11.93
13	11.97	11.95	11.92
14.7456	11.95	11.93	11.91
15.36	11.95	11.92	11.9
16	11.94	11.91	11.9
18.9583	11.92	11.87	11.85
19	11.92	11.87	11.84
19.2	11.91	11.86	11.83
19.9899	11.91	11.85	11.83
20	11.9	11.85	11.83
20.48	11.89	11.84	11.8
23.04	11.85	11.8	11.77
40	11.77	11.61	11.57

Table 5.
Measured power level across temperature.
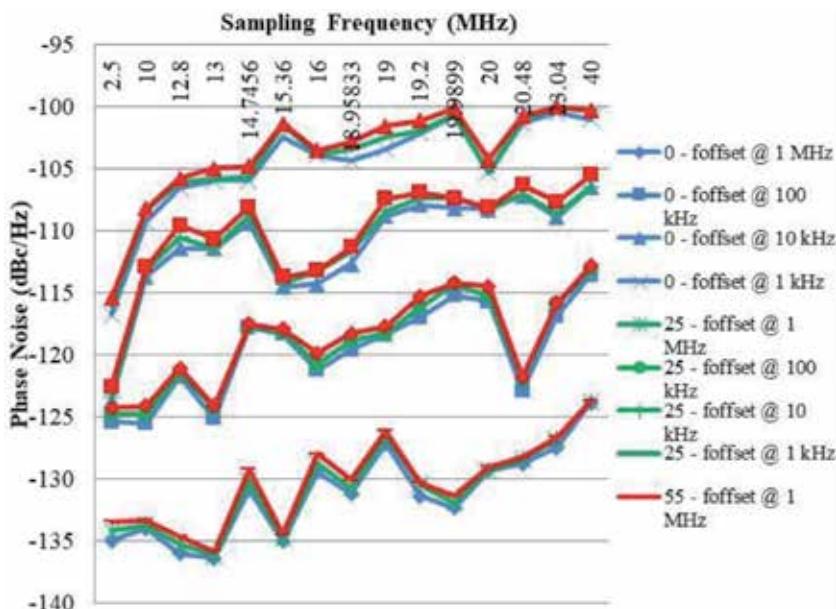
Figure 12.
Phase noise result over temperature.

There are four parameters involved in jitter analysis. Those are random jitter (RJ), period jitter (PJ), data-dependent jitter (DDJ), and total jitter (TJ) (**Figure 14**). In the second graphs of **Figure 14**, the RJ and PJ have been zoomed in. This graph is important to this work as it is a measure of how stable the clock is.

Apparently the poorest jitter fell upon a 20.48 MHz WCDMA sampling clock. The cause could be contributed by the DDS internal logic itself. Even with a
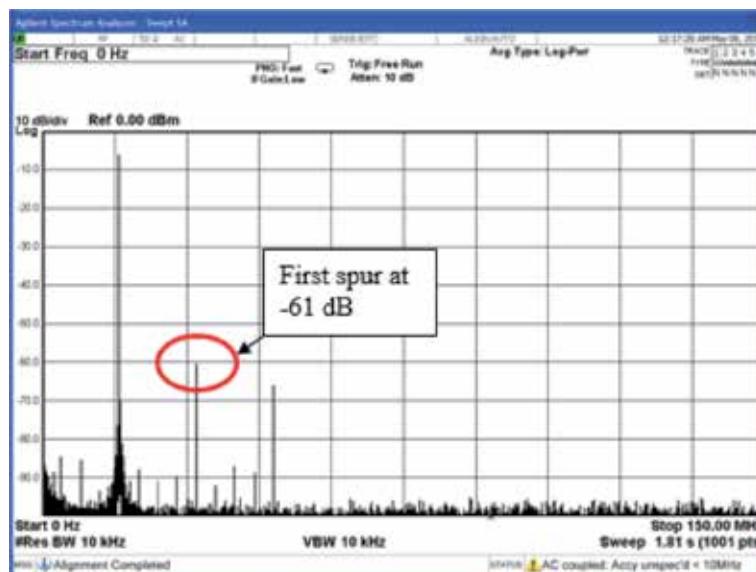
Figure 13.
SFDR at frequency 16 MHz.
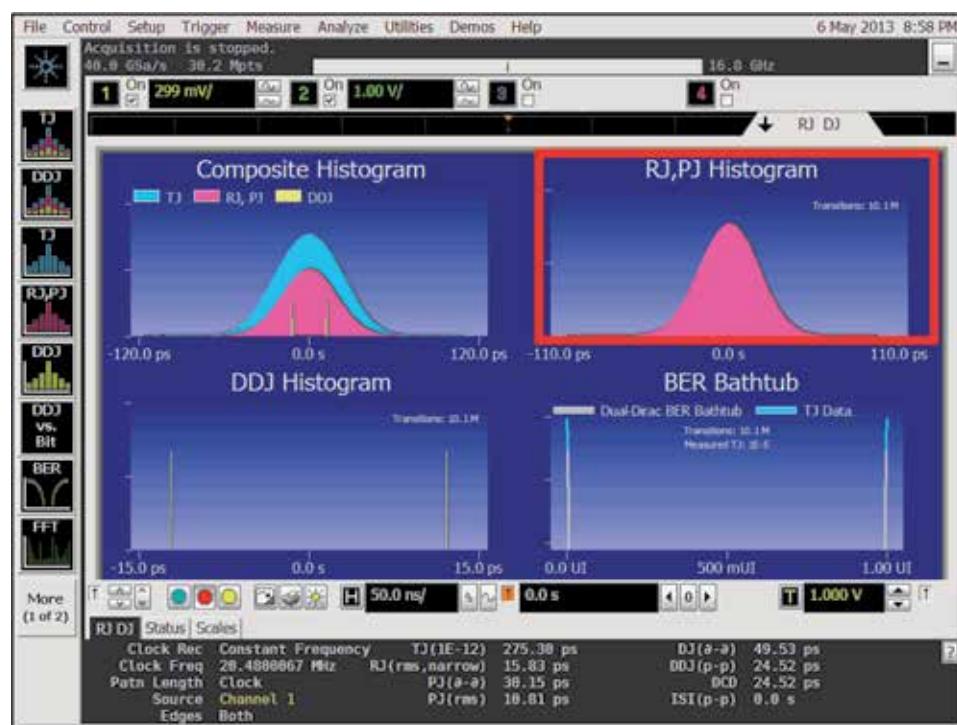
Figure 14.
Output clock jitter at 20.48 MHz.

30.15 ps, its output is far better than the DMP PLL that has a peak jitter at 150 ps. The same goes to jitter RMS. When the output frequency is at 20.48 MHz, the RMS jitter was measured at 10.81 ps. The accumulative transition of 10.1 M samples yields that the output clock at 20.48 MHz is fluctuating around the time axis. The

ADC chip has a requirement of 200 ps which makes the DDS meets the criteria. Otherwise, the output from the ADC will be subjected to nonlinearity errors.

The RMS period jitter is determined using the standard deviation of random jitter. As far as the random jitter is concerned, it is composed by the random effect of thermal, flicker, and shot noise that inherits within the DDS. Those noises are a function of temperature that will typically degrade the period jitter performance. However, in this work, the DDS has a remarkable output that remains within 12 ps PJ$_{RMS}$ at 55°C (refer to **Figure 15**).

Both terminations with high and low impedance are performed in transient analysis. From the results as shown in **Figure 16**, it is obvious that when the measurement was made using high-impedance probe, there is a reflection which occurred at high state (3.3 V) as well as low state (0 V). The voltage overshoot can
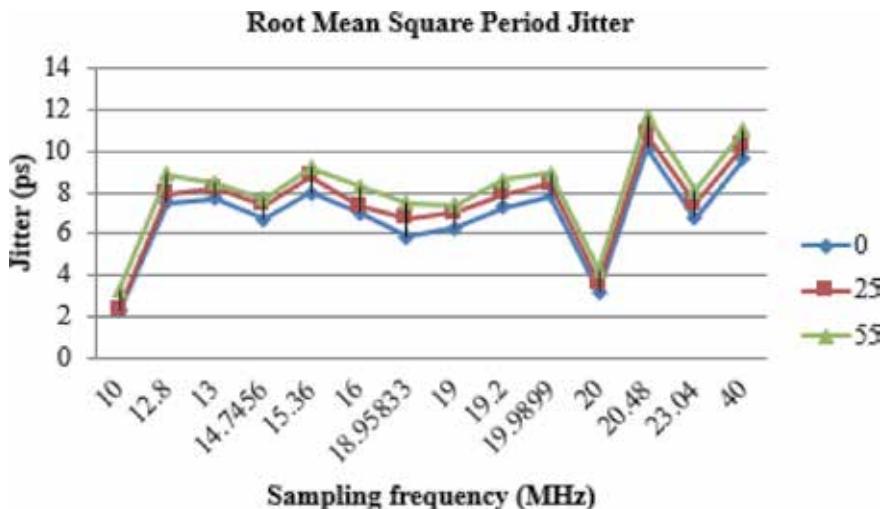
Figure 15.
RMS period jitter across temperature.

Figure 16.
Clock output waveforms comparison at 2.5 MHz.

go as high as 4.2 V and as low as 0.4 V. At low frequency of 2.5 MHz, the overshoot period is shortened and settled almost immediately. However, as frequency increased, the overshoots stay longer to settle. This is essential in order not to damage the subsequent device.

One possible solution is to connect this output to a buffer before sending it into the next stage. Alternatively, a matched load to this output will eliminate the overshoots as shown using low-impedance probe of 50 Ω. By matching the comparator output to a 50 Ω, a clean 2 V_{p-p} signal can be achieved. In any case, the final implementation utilizes a 50 Ω balun to generate an unbalanced output, and therefore, differential signal is obtained. The balanced to unbalanced signal conversion may generate ringing around two different states, but it also provides an isolation between the source and the receiver in this case, DDS output and ADC sampling clock input. Though there are ringing at both differential signals, the differential ADC clock input only accounts for the difference between two signals to define as a transition.

3.3 Comparative result analysis

The filter design has step forward from the targeted specification into the simulation and then the actual implementations. By catering the worst-case analysis through simulation, the final outcome has somewhat stratified. The measured value was retrieved from an average of five samples at ambient condition. The targeted cutoff frequency was at 42 MHz as to allow a wider design margin for the filter to shift this frequency around. The implemented design shows a 41.95 MHz unlike what the simulation preached at around 46 MHz. When subjected to temperature, the poorest corner frequency fell at 41 MHz with 1 MHz margin to designed frequency. **Table 6** tabulates the filter comparison.

Secondly, the gain at 40 MHz is relatively low compared to the simulation result. Actual LPF exhibits a −2.23 dB loss compared to targeted 0 dB. However, this loss is essential to the design in this work. Since a high-speed comparator that supports hysteresis of maximally 20 mV_{p-p} is being used, the outputs from AD9852A has been programmed as differential; these −2.23 dB loss seen at the output of LPFs will become insignificant. Though with only 77.4% strength of the original signal, as long as the output is differenced by 20 mV_{p-p}, the comparator can still produce a clean and stable clock signal.

For the stopband gain, apparently the actual measurement shows a better performance than the simulated result from the ADS. The difference between MATLAB and design guide is anticipated. The ideal value to nominal value conversion will definitely contribute that loss into the design. Of course it is far worst compared to the targeted value. As a reminder, the targeted value was buffered by 8 dB, and the original specification is 70 dB. In this case, the actual design is better in essence. Lastly, the phase response for all results yells the same thing. The designed filter has a linear phase across the passband.

Criteria	Target	MATLAB	ADS design guide	Final	Measured (25°C)
3 dB cutoff ($f_{cut-off}$)	42 MHz	46.2 MHz	46 MHz	46.06 MHz	41.95 MHz
Gain at 40 MHz	0 dB	−0.006 dB	−0.004 dB	−0.003 dB	−2.23 dB
Stopband gain (A_{stop})	−78 dB	−78 dB	−78.07 dB	−71.561 dB	−72.66 dB
Phase response	Linear	Linear	Linear	Linear	Linear

Table 6.
Filter specification comparison.

Sampling rate (MHz)	Simulation	Measured (25°C)	% diff.
2.5	11.895	10.39	14.49%
10	11.876	11.95	0.62%
12.8	11.869	11.95	0.68%
13	11.865	11.95	0.71%
14.7456	11.858	11.93	0.60%
15.36	11.855	11.92	0.55%
16	11.854	11.91	0.47%
18.9583	11.839	11.87	0.26%
19	11.838	11.87	0.27%
19.2	11.837	11.86	0.19%
19.9899	11.831	11.85	0.16%
20	11.833	11.85	0.14%
20.48	11.828	11.84	0.10%
23.04	11.811	11.8	0.09%
40	11.529	11.61	0.70%

Table 7.
Power level result comparison.

The comparison was also made in terms of spectrum analysis. Both results collected based on the 50 Ω termination are shown in **Table 7**. Generally, there are some differences from a frequency range of 10 to 40 MHz. The discrepancy occurred to be lesser than 1%. The only largest difference in magnitude is found at the sampling frequency of 2.5 MHz. The loss could be imposed by the PXA itself or even the bandwidth of the comparator itself. However, given the advantages of having balun 50 Ω matched to the comparator and a differential clock requirement for the ADC device, this loss will become insignificant.

Specifications	DDS	DMP PLL	Improvement
Frequency range	DC-42 MHz	2.5–40 MHz	12.0%
Frequency resolution	1.1 µHz	3.73 mHz	Huge
Phase noise at 1 kHz	<100 dBc/Hz	< 80 dBc/Hz	25.0%
Period jitter (rms)	12 ps	50 ps	76.0%
Period jitter (peak-peak)	121 ps	300 ps	59.7%
SFDR	70 dBc	56 dBc	25.0%
Tuning speed	<100 µs	< 4.3 ms	97.7%
Communication protocol	I^2C	SPI	—
Communication speed	2.8 MHz	1 MHz	180.0%
Number of wire	2	3	33.33%
Part counts	40	183	78.1%
Cost	$19.88	$64.90	69.4%
Size	50.8 × 38.1 mm	76.2 × 50.8 mm	50.0%

Table 8.
Comparison between DDS and DMP PLL.

Table 8 summarizes the improvements between DDS and DMP PLL. The new DDS has taken all consideration in order to ensure that the fundamental performance of DMP PLL remained without degrading the system performance. Besides, huge improvements were achieved. The DMP PLL has an operating range from 2.5 to 40 MHz, while the DDS can go as low as 1.1 μHz where it has been declared as DC. A total of 12% improvement has been achieved. On the other hand, the frequency resolution improved significantly as compared to DMP PLL.

For phase noise, the DMP PLL has a − 80 dBc/Hz, while DDS offers 25% lower than −100 dBc/Hz across temperature. With this achievement, both peak-to-peak and RMS period jitters have been improved by 59.7 and 76%, respectively. This will result in the error and improvement measurement speed by reducing the cyclic redundancy check (CRC). For SFDR, DDS offers 25% of improvement with 70 dBc as compared to DMP PLL with 56 dBc. In the tuning speed, a remarkable 97.7% speed improvement has been achieved by the improvement of communication speed with DDS. This number can go below if future development decides to increase the speed further for faster measurement. Lastly, the part counts, size, and cost were interrelated. The reduction of part counts reduces the space required and, therefore, the cost reduced.

4. Conclusion

By having every piece a part of the design, the DDS core together with the LPF in the design is connected together to perform as a clock source for ADC sampled system. Both simulation and measurement have demonstrated almost identical performance for various sampling frequency ranges across several temperature setups. Overall, the design has achieved a peak jitter of 60.5 ps and RMS of 12 ps across all sampling frequencies. Phase noise has been improved by 20 dB below the original design for a −100 dBc/Hz at 1 kHz frequency offset. Tuning speed has fastened from 4.3 ms down to 100 μs with improvement over communication speed. Total part count has been reduced by 78.1%, and the circuit space shrunk by 50%.

Author details

Desmond Tung and Rosmiwati Mohd-Mokhtar*
School of Electrical and Electronic Engineering, Universiti Sains Malaysia,
Nibong Tebal, Pulau Pinang, Malaysia

*Address all correspondence to: eerosmiwati@usm.my

IntechOpen

References

[1] Parker M. Automotive Radar. In: Digital Signal Processing 101. Oxford, UK: Newnes; 2017. pp. 253-276

[2] Munoz-Ferreras JM, Gomez-Garcia R, Li C. Human-aware localization using linear-frequency-modulated continuous-wave radars. In: Li C et al., editors. Principles and Applications of RF/Microwave in Healthcare and Biosensing. London, UK: Academic Press; 2017. pp. 191-242

[3] Peyton AJ. Electromagnetic induction tomography. In: Wang M, editor. Industrial Tomography: Systems and Applications. Cambridge, UK: Woodhead Publishing; 2015. pp. 61-107

[4] Hughes L. Direct digital synthesis vs. direct analog synthesis: The future of frequency agility. 8 May 2017. [Accessed: 29 Aug 2019]. Available from: https://www.arrow.com/en/researchand-events/articles/dds-vs-das

[5] Murphy E, Slattery C. Direct digital synthesis controls waveforms in test, measurement and communications. Analog Dialogue. 2005;**39**(8):1-4

[6] Chao H, Li-xiang R, Er-ke M. A systematic frequency planning method in direct digital synthesizer (DDS) design. In: International Conference on Wireless Communications & Signal Processing. 2009. pp. 1-4

[7] Xiaoming D, Jun Z. DDS phase-locked swept source and study design. In: International Conference on Computer Science and Network Technology (ICCSNT). 2011. pp. 146-149

[8] Nosaka H, Yamaguchi Y, Yamagishi A, Fukuyama H, Muraguchi M. A low-power direct digital synthesizer using a self-adjusting phase-interpolation technique. IEEE Journal of Solid-State Circuits. 2001; **36**(8):1281-1285

[9] Avitabile G, Cannone F, Vania A. Phase shifter based on DDS-driven offset-PLL. Electronics Letters Journal & Magazine. 2006;**42**(25):1438-1439

[10] Young-Wan K, Jae-Du Y. Phase noise model of single loop frequency synthesizer. IEEE Transactions on Broadcasting. 2008;**54**(1):112-119

[11] Cetin E, Kuyel T. Hundred femtosecond jitter measurement using dual-channel ADC with imprecise clock stimulus. Electronics Letter Journal & Magazines. 2012;**48**(24):1525-1526

[12] Jinwook J, Franza O, Burleson W. Period jitter estimation in global clock trees. In: 12th IEEE Workshop on Signal Propagation on Interconnects. 2008. pp. 1-4

[13] Analog Devices, AD9852 CMOS 300 MSPS Complete DDS, Technical Datasheet. 2008. pp. 1-41

[14] Hee-Ran A, Tatsuo I. Isolation circuit of impedance-transforming 3-dB compact baluns for near perfect output matching and isolation. IEEE International Microwave Symposium Digest. 2010:113-116

Understanding of On-Chip Power Supply Noise: Suppression Methodologies and Challenges

Pritam Bhattacharjee, Prerna Rana and Alak Majumder

Abstract

The on-chip activities of any modern IC are always inhibited due to the occurrence of power supply noise (PSN) in the chip power line. From many decades, researchers are pondering on what are the major issue of this PSN occurrence and how it can be suppressed without interfering the actual chip functioning. In the course of time, it is found that the uncontrolled triggering of the on-chip system clock and the unguarded on-chip power line is instigating the two major factors for the occurrence of PSN i.e., i(t) → instantaneous current and di/dt → current ramp or the rate of change of current over time. Both i(t) and di/dt are also the sub-factors to rise the PSN components like resistive noise and inductive noise respectively. In this chapter, we light upon the occurrence of resistive and inductive noise as well as depict their individual impact on the PSN occurrences. There is also discussion on how PSN is suppressed over the years in spite of facing challenges in the execution of suppression techniques. This chapter even concludes on the suitable ways for mitigating PSN in the contemporary era of delivering complex on-chip features.

Keywords: power supply noise, resistive noise, inductive noise, variable frequency clock, clock gating

1. Introduction

The business of integrated circuit (IC) chip has become an all-time growing venture since the last few decades. In order to continue the legacy, there are always changes and developments in the method of traditional chip design, addressing lot of issues related to the proper functioning of circuits inside the IC. The chip designers and the researchers constantly explore deep into these problem areas to find out their appropriate solution. The basic functioning of all inside the chip circuitries is primarily dependent on the power supply (V_{dd} and ground) lines attached to it. However, these V_{dd} and ground lines are supplied externally through metal pins of the outer package. Not only the proper power supply lines help the circuit performance inside an IC, but also it supports the logic level and signal integrity of all incoming and outgoing signals on and off-chip.

However in actuality, any signal entering the outer package through the metal pin has to experience the current/voltage level deterioration due to the presence of package parasitic resistive and inductive components along the pathway of metal pin, bond wire and bond pads of the IC chip as seen from **Figure 1**. Typically, the external metal pins that are soldered to the bond pads of an IC chip, have improper soldering, create
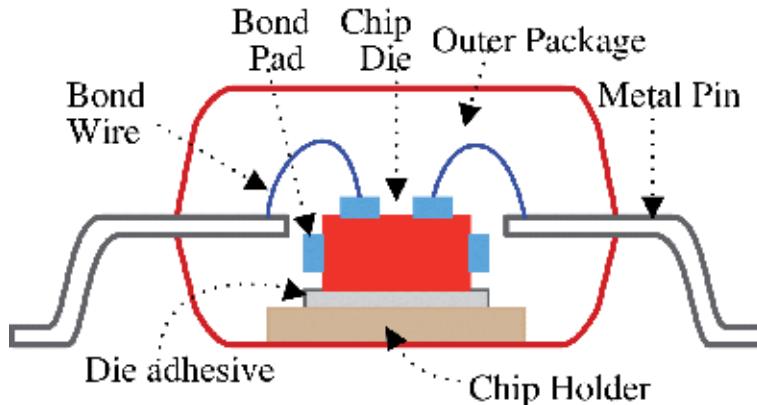
Figure 1.
Cross-sectional view of an IC chip.

tiny gap between the chip and the chip holder. For this reason, insulated adhesives are used to fill up the gaps so as to maintain safe signal transmission. In fact, these insulating adhesives always relieve the soldered joints by providing them strong mechanical connection. But post-manufacturing of the chip, these adhesives wear out with time and there is issue with the appropriateness of signal integrity inside the chip and this is one of the prime reasons for the instigation of power supply noise (PSN) in integrated circuits. The problem of PSN is more accurately stated as the fluctuation in the power supply signals, correlated to the rate of flow of current in any circuit (inside chip) under the influence of resistive and inductive parasitic attached to the circuit PDN (power-delivery-network). Sometimes, the PSN is also referred as switching noise which is the primary source of fluctuation in power supplies and is generally quantified by the maximum droop or peak in the supply voltage with respect to the corresponding nominal value [1]. Even the peak-to-peak supply voltage and the average value of supply voltage are quoted as the measure for PSN [2]. It is to be noted that the critical path delay of logic circuits inside the IC chip is inversely proportional to the rate of flow of supply voltage [3]. Any small-scale drop in the supply voltage induces timing delay for the functioning of logic circuits and if the value of delay is high in magnitude, it can even impose logical error to the operation of IC. For example, if we consider any general clock tree network (distributing clock signal to any of the sequential circuits inside the chip), there ought to be delay in the clock line while there is drop in V_{dd}. Since the clock signal from that clock tree is delayed, there happens to be clock skew with respect to the same clock signal which is transmitting in other lines. Besides, those sequential circuits inside the chip face setup time violation due to the inappropriate arrival of clock, hence always culminating logical error in their functionality. Not only it is the delay which is badly getting impacted, but also the occurrence of PSN drops the performance consistency of many circuits inside the chip as they are sensitive to the requirement of minimum supply voltage [4]. Therefore, it is very important to understand the issues of on-chip PSN resolve them to an efficient level and take a note of what the challenges are along the pathway of solving the issues.

The book chapter is organized as follows: Section 2 and its sub-sections will contain discourse of the exact problems depicted by PSN in IC chips. In Section 3, the on-chip PDN (power-delivery-network) has been discussed which actually distributes the power supply to circuits inside the chip and is one of the major victims of PSN. We talk about the primitive methodologies of suppressing on-chip PSN in Section 4 and highlight the challenges faced during this activity in Section 5. Lastly in Section 6, we have tried to put in our part of solution to the existing PSN problems and end with our concluding remarks.

2. State of art for the on-chip power supply noise problem

As a matter of fact, the amount of PSN along the circuit power line inside the chip is fundamentally the function of switching activity in the circuital nodes and consequently the instantaneous current pumped inside the system. This current pumping happens through the PDN inside the chip and cause voltage drops in the circuit power lines due to the presence of inductive, capacitive and resistive elements in the power distribution grid of the IC chip. The nominal V_{dd} (supplied from the off-chip space) get cut down to some degraded level (as shown in **Figure 2**) by the time it reaches the circuit power line.

Now, such occurrences are prevalent because of two major factors which are as follows: (i) the circuit current pumping is high with certain dip in the nominal V_{dd} and V_{gnd} because of the presence of parasitic resistance and it is often referred as "IR noise" or the "resistive noise", which is denoted by ΔV_R. The relationship of ΔV_R has been depicted in Eq. (1).

$$\Delta V_R = i(t) \times R \qquad (1)$$

where, R is the parasitic resistance and i(t) is the instantaneous current. (ii) The rate of change of current across the parasitic inductance resulting "di/dt noise" or the "inductive noise", denoted as ΔV_L. The relationship of ΔV_L has been depicted in Eq. (2).

$$\Delta V_L = L \times \frac{di(t)}{dt} \qquad (2)$$

where, L is the parasitic inductance and the rest is current ramp (di/dt) over a certain period, inducing fluctuations to the supply voltage level (specially to the V_{dd}).

In virtue of the relationships stated in Eqs. (1) and (2), it is inferred that higher the logic level switching across the circuital nodes within an IC chip, more is the current drawn in presence of the inductive and resistive effects which result in large PSN along the circuit power lines. This is because PSN happens because of the combined impact of inductive and resistive noise, depicted as $\Delta V_R + \Delta V_L$ [5]. As a matter of fact, PSN is not only a major concern for the mixed-signal systems or the analog circuits, but also it adversely interferes the logic and timing performance of pure digital circuits [6] present inside the chip. The analog circuits which comprise of sensitive bias points fluctuate due to the variation in supply voltages and hence are very prone to the adversities of PSN. It affects important performance circuit parameters like current/voltage gain and defects the linear behavior of most of these circuits. In case of digital circuits, though the adversity due to PSN
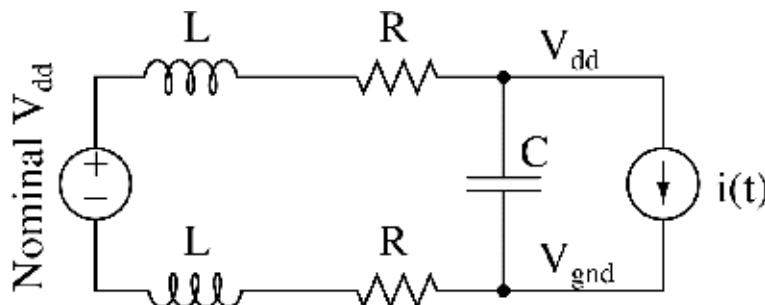
Figure 2.
Simplified circuital scheme of power distribution system.

is less, but there is always a genuine possibility that the noise level is more than the circuit noise margin specifications. Therefore, it is very important for circuits (be it analog, digital or mixed-signal) which reside inside the chip, to have guard against the PSN taint because the consequences are extremely fatal. For example, the critical path delay of circuits is inversely proportional to the amount of supply voltage. In fact, with technology scaling, the gate delay is even more sensitive to the supply voltage variations [7]. A case study which has been reported in [8] depict that 1% of supply voltage variation is causing approximately 4% variation in the delay parameters. In fact, inside the chip storage elements like latches, flip-flops, registers and SRAMs (static random-access-memory) have strict limitation of having a minimum supply of voltage (let us say, V_{min}) which determines the logic level of the stored data. Any sort of deviation in the required value of V_{min}, can lead to error in the stored data.

Another nasty consequence of PSN is that it adversely impacts the timing signals of digital clock propagating inside the chip due to the intrusion of timing jitter. Basically, the digital clock is distributed inside the chip through clock distribution network (CDN) comprising of dense clock trees that are connected to several sequential blocks inside the chip. These clock trees are made of cascaded buffers arranged in various ways based on which it is typecast as either balanced or unbalanced tree or other types [9]. The power supply (V_{dd} and ground) of these individual buffers are connected to the chip power line as shown in **Figure 3** through different tapping points and the buffers draw current while the clock signal is propagating. Now typically, V_{dd} pins of all these individual buffers are never tapped to single tap point of chip power line so as to make sure that the circuit loading is moderate at any particular point.

Now, when the PSN occurrences contaminate the chip power line, it is obvious that there is voltage fluctuation in V_{dd} across each tapping points, however with different intensity. Therefore, it is very likely that the propagation delay of clock signal through different tapping point will have different timing even if the
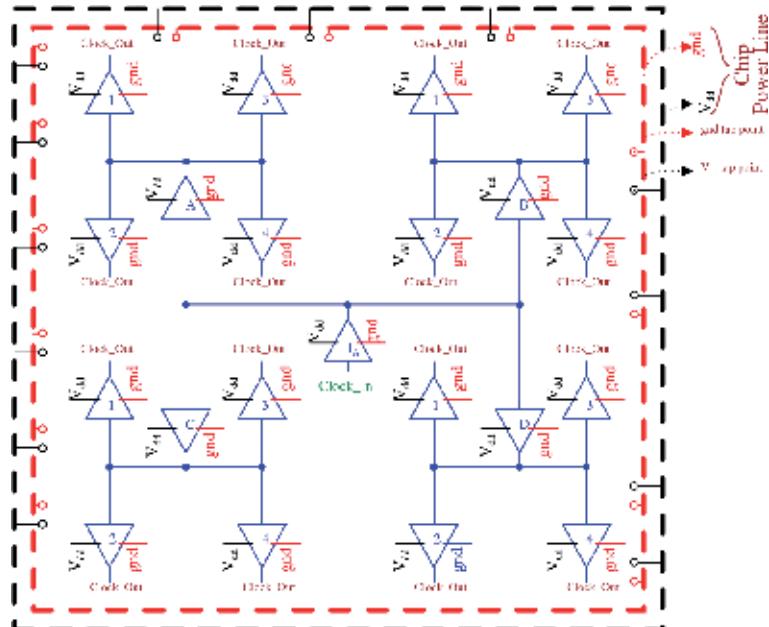
Figure 3.
H-type balanced clock tree network as a part of CDN.

buffers are identical. For example, the clock inserted in the buffer "I_A" outputs the clock signal through the output node of identical buffers "A_1"..."A_4", "B_1"..."B_4", "C_1"..."C_4" and "D_1"..."D_4". But as the V_{dd}-level of these buffers is different due to PSN, their "clock_out" has different pitch time on the sequential blocks to which they are connected. In fact, most of the "clock_out" are considered as "faulty" as per their individual timing, causing clock synchronization error to the operation of sequential blocks. Apart from these types of issues, there also remain anxieties with the circuit performance reliability when the supply voltage is high compared to its expected value. Certain overshoot in the supply voltage even cause electromigration problems [10] and hot carrier effects (HCEs) [11] in the circuit design and disturb its actual functionality. Therefore, the occurrence of PSN is always an apprehension for the circuit designers.

The mitigation of PSN is generally executed by monitoring the voltage-level fluctuations across circuit power line either continuously over time (i.e., active mitigation) or at certain instances in time (i.e., passive mitigation). However, it is also important to get hold of how much supply noise is occurring, i.e., detection of on-chip PSN. For that, the understanding of on-chip power-delivery-network (PDN) is essential which is being discussed in the upcoming section.

3. Discussion on the power delivery network for the chip

There are several ways using which the chip is facilitated by the power line supplies. Basically, the crucial aspect in the design of power delivery network (PDN) is the dedicated routing styles, for example, the ad-hoc style, the grid-like structures or even the power and ground planes. In the ad-hoc style, local circuit blocks are connected to the I/O pads (on the outer boundaries) using customized power/ground lines as shown in **Figure 4(a)**. On the other hand, grid-like PDN approaches which are commonly found in modern high-power ICs [12] is projected in **Figure 4(b)** where generally multiple layers are used in the design of power grid and power/ground lines are interdigitated within each layer and the layers are orthogonal to each other. The reason why these grid-like structures are popular in modern ICs is that the area between the power and ground line has lot of unused place which has the scope to be utilized for signal routing. This makes the power supply level insensitive to the current prerequisites of various circuit blocks. Therefore, some failure in power delivery to any one of the blocks does not prevent flow of power supply to the remaining circuit blocks. In fact, the grid structure supports the possibility of implementing dual type power supply voltages, which is essentially required in high-performance chips so that power consumption can be reduced [13]. The grid structures also facilitate the option for individual supply lines be connected to the appropriate supply voltages.

In fact, there is another design approach of PDN depicted in **Figure 4(c)**, however it is not competent in contrast to the other two approaches even though it
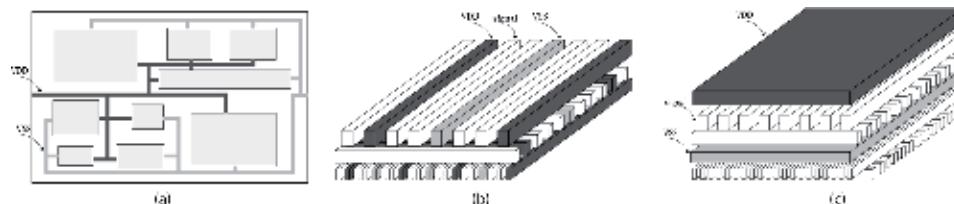
Figure 4.
On-chip PDN styles viz., (a) ad-hoc style (b) grid-like structure and (c) power and ground planes.

provides low impedance path for the current flow. The major setback here is that there is no available space for rigorous signal routing [14]. So, in the recent days the grid-like structure for PDN design is pursued for all types of on-chip architectures.

But it is important to remember while experimenting with the various PDN approaches that the grid impedance should not exceed the targeted limit which is stated in Eq. (3).

$$Z_{target} = \frac{V_{dd} \times ripple}{I_{max}} \tag{3}$$

Here, V_{dd} is the nominal supply voltage delivered from the off-chip area and I_{max} is the maximum on-chip current pumped while the full chip functionality is in execution. In this context, the term "ripple" illustrates the amount of noise (i.e., level of voltage fluctuation/ripple) allowable on the power supply line. Typically, for the case of silicon chip having 1 V power supply with the maximum current pumping of 100 A, the maximum allowable ripple is 10% because of which the Z_{target} is alarmingly low i.e., 1 mΩ. However, there is an important point to be taken care that the Z_{target} at each level of PDN must be relevant across the current transient times [15] or else there is the possibility of PSN contamination. So, in most ideal cases, it is observed that the on-chip current transient times are around 10% of the on-chip clock frequency.

4. Methodologies to suppress on-chip power supply noise

The impact of PSN over on-chip circuits is suppressed using different reduction and isolation techniques. These strategies are broadly classified in three areas as depicted in **Figure 5**.

The very fundamental method of stopping the generation of PSN is to increase the current transient time because that subsequently reduces the di/dt noise. In fact, there is a similar approach which is known as current spreading [16]. This technique spreads out the switching activity over an extended period and thereby decreases the amount of i(t) pumped into the on-chip circuits.

4.1 Circuital approaches

One of the primitive circuital attempts to increase the current transient time is reported in [17]. Stretching the time stamp actually helps to compute the average value of i(t) at moderate level. This helps in dropping the content of resistive noise as well as some part of inductive noise since the value of current ramp – di/dt is also at low range. Hence cumulatively, PSN is diminished. Even the noise immunity to PSN is increased with the incorporation of circuital tweaking [6]. For instance, the weak voltage keepers can be used in the design of dynamic logic based on-chip circuits which can significantly reduce the effects of noise i.e., eliminate the logical failures in these circuits by providing weak feedback to their critical circuital nodes. In case of analog circuits, differential signaling is used to validate the occurrence of PSN as common-mode noise [6]. Since the differential signaling creates an on-chip communication between digital and analog modules inside the chip to have separate individual power supplies from the main supply source, there is least chance that the digital V_{dd} intersecting the analog one. So there is voltage of V_{dd} remain intact. However, in situations where the circuital approaches are not enough to alleviate PSN, people try to incorporate the non-circuital approaches.
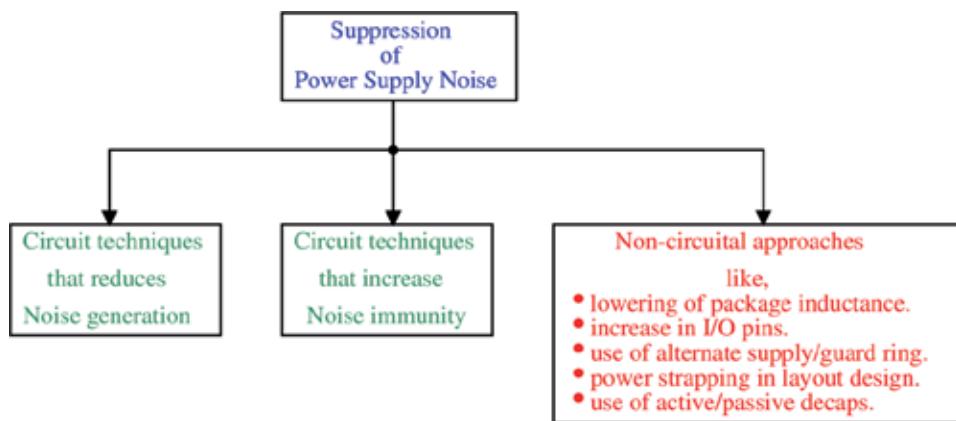
Figure 5.
Various strategies for PSN suppression.

4.2 Non-circuital approaches

Basically, the non-circuital approaches are the ones which can be imposed only during the chip fabrication. Like for example, the level of parasitic inductance in the supply path is lowered by implementing C4 flip-chip technology instead of typical wire-bonding technology. This helps in the reduction of inductive noise and even the resonance effect in any system. Consequently, the occurrence of PSN is suppressed to a substantial level. Another strategical approach is to increase the number of pins assigned for the power supply section. However, in this particular approach there is always physical limitation like the pin count can be predefined. So typically, alternative and multiple power supply sources are used for critical on-chip circuitries, like, the analog domain and digital domain has dedicated and independent power supplies such that the analog circuitries can never get hindered by the noisy digital supplies. Besides, there is always some guard rings to shield the power supplies from the contamination of noise. But the efficacy of these guard rings in most modern process technologies is questionable due to the use of low resistive substrates [18]. Therefore, the physical designers are more enthusiastic to play with the routing schemes of the power supply section which is often referred as power strapping [19]. Apart from this, one of the popular non-circuital approach to suppress PSN is by implementing on-chip decoupling capacitors (abbrev. "decap") which are of two different types viz., active decap and passive decap. The placement of small decaps in contrast to single large device reduces the impact of parasitic resistance and inductance in the supply line [20]. Hence, the rate of PSN is diminished. Typically, the passive decaps are preferred over the active decaps because the rate of noise suppression is higher in case of passive decaps compared to the active decaps [20].

As a matter of fact, the non-circuital approaches are the fixed solution for the alleviation of PSN, but their execution process and engineering is very costly. From the circuit designer point of view, we would prefer circuital tweaking and techniques to reduce PSN, however accepting that there are also challenges.

5. Challenges faced in suppressing on-chip power supply noise

In the earlier sections, there has been discussion on problems asserted by PSN to the proper functioning of on-chip circuitries inside the modern ICs. The typical PSN suppression methodologies was highlighted. But reducing PSN can be very

challenging. For example, the circuital modifications which are planned to be implemented for reducing PSN are rarely process independent. So generally, the extra transistors that are employed for checking the level of PSN, behave different from what is expected and the suppression of PSN is hindered. Another major issue is the scale down of process technology in the recent days. The impact of technology scaling over signal-to-noise ratio (SNR) with respect to the on-chip clock frequency is really adverse because the SNR is never constant in presence of fluctuating clock frequency due to scaling.

It has been mentioned about the use of weak voltage keepers for suppressing PSN. But in actuality, the weak feedback provided by these keepers does not prevent delay failures and hence it cannot be inferred as the general-purpose solution for noise suppression. Even the use of different V_{dd} for analog and digital domain faces problem like the voltage headroom (i.e., voltage range within the V_{dd} and ground plane) is different in each case and wherever the on-chip analog/digital circuits are interfaced, most of the time, there is issues with logic level of the propagating signals. In fact, the differential signaling used by the analog circuits to reduce PSN is only possible if both the analog/digital section share a common ground line. But as per the chip requirement, the analog ground is mostly kept separate from the digital ground so as to avoid resonance. Therefore, the use of differential signaling is conditional and rarely approached for the suppression of PSN.

However, there has been lot expectation with the implementation of decaps in context to the suppression of PSN, but it also faces crucial challenge while doing the job. Basically, the decap absorb supply noise like there happens to be drop in battery potential while the battery is connected to any electrical load. But again, similar to the battery, the decap requires definite time for charging and discharging which oppose the quick changes in supply voltage. In fact, the on-chip reactance of these decaps is so high during low frequency functioning of the chip that as soon as the supply voltage fluctuates, it badly affects the on-chip performance. Therefore, the PSN suppression using decaps is rarely accepted worldwide in the recent days.

In retrospective of the present scenario on how suppression of PSN is facing challenges, both the circuit designers and researchers are looking for appropriate solutions which can mitigate PSN irrespective of any dependencies. From the last decade, there has been some potential circuital approaches coming up which has shown good result in the suppression as well as in the alleviation of PSN.

6. Probable solution to reduce the on-chip power supply noise

An attractive solution to reduce the PSN has been found in recent years with the incorporation of variable frequency signaling (i.e., some MHz to GHz as shown in **Figure 6**) because it has the nice ability to curb random fluctuations in the power supply voltage or current [21].

The variable frequency clock (VFC) as it is popularly known, is used in the timing and control unit of the modern processors to automatically adjust the frequencies of various components inside the processor [22–24]. The measure of frequency directly impacts the switching activity of the on-chip circuital nodes. So, when the frequency is low, and the corresponding current transient time is high, the amount of i(t) and di/dt is less. At a stretch, as the frequency increases, the switching activity increases along with the measure of i(t) and di/dt. But in the average, both i(t) and di/dt has moderate values which is the main reason why PSN get reduced when the conventional system clock (typically in the range of some GHz) is replaced with VFC inside the chip.

Basically, the VFC is inserted to CDN and connect the corresponding clock trees. So, when it is propagating through the lines (like the way shown in **Figure 3**), the buffers draw current from the power supply (V_{dd} and ground). But this time, the buffer input is a VFC rather than a conventional clock, so the average value of voltage fluctuation (which occurs due to PSN) is very nominal and does not impact the timing parameters. Therefore, the inclusion of VFC instead of conventional clock helps the CDN to synchronize the clock signal at each and every sequential block inside the chip.

Apart from this advantage, the inclusion of VFC also help to modulate the flow of i(t) for any inside the chip sequential circuit. Typically, the i(t) plot is similar to what is shown in **Figure 7(a)**. But since the current pumping due to VFC in the initial stages till wake-up is slow and steady, the i(t) plot ideally looks like the one shown in **Figure 7(b)** and after the VFC reach its highest frequency, the current profile is almost similar to that of conventional clocking. This is also an added advantage for the average di/dt and reduces the impact of PSN.

However, the initial designs of VFC could not generate right pattern of clock (like the one shown in **Figure 6**) because of which the values of di/dt as well as the frequency ramp (i.e., change of frequency per unit time) has been random and large, leading to no control over PSN. In 2010, the IBM (International Business Machines Corporation) T.J. Watson Research Center [25] reported a SOI (System-On-Insulator) based VFC which produce frequency starting from 800 MHz to 12GHz with a minimum average frequency ramp of 17.5 MHz/μs and impressively reduced the measure of di/dt as well as PSN. But, the architecture of this VFC comprise of complex circuitries like digital phase locked loop (DPLL) and digitally controlled oscillator (DCO) which are power-hungry designs and derive large instantaneous current over a period of time. In fact, in 2018 also, a group of HKUST – Hong Kong [26] came up with a VFC design (made of time-to-digital convertors and digital loop filters) to generate variable frequencies having tuning range from 82GHz to 108GHz stirring a high start-up current with its peak value being quite high. Though, both these VFC design could successfully reduce PSN, their architectural complexities created scope for other researchers to find out more appropriate solution.

Hence, the work of IBM researchers [25] was carried forward in [27] and their VFC architecture was redesigned to develop a tuning range from 250 MHz to 2GHz with only the involvement of components like flip-flop-based frequency dividers, multiplexers and simple decision circuits. In fact, while checking the performance of this VFC to drive a sequential core, it was found that the peak position of the current drawn inside the core is attained after 40 clock cycles, which is otherwise noted only after 10 clock cycles when steered by the conventional system clock. Thereby, this VFC architecture ascertains long wake-up time for the system clock, thus reducing the i(t) and di/dt. But this VFC failed to attain the current transient time as small as compared to what was achieved by the IBM researchers. So, this VFC architecture was modified by reducing the circuit overhead and distributing the modules and sub-modules of the VFC in different cores of the chip [28]. That helped outstandingly in the improvement of i(t) and di/dt and also reduced the PSN with much effect.

However, as it is noted that the switching activity is a prime facet for the instigation of i(t) and di/dt, we have tried implementing clock gating to the conventional
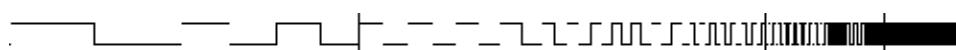
Figure 6.
Example of a variable frequency signaling.
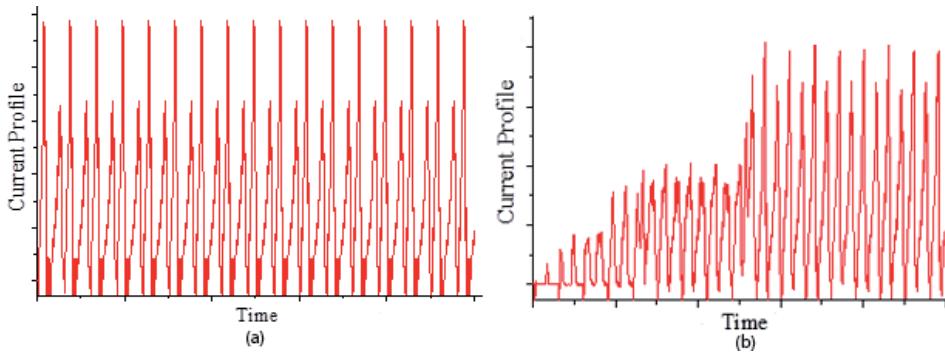
Figure 7.
Instantaneous current flow due to (a) conventional clock (b) VFC.

system clock so as to restrict the activity factor (α). There are many on-chip components which do not function at the same time instance like all the others. But because they are connected to the same system clock, the current pumping (i.e., i(t) and di/dt) happens across these components in spite of they are not participating to the on-chip activity at that point of time. In [29], we have already projected how the implementation of clock gating has helped in the minimization i(t) and di/dt as well as the curbing of PSN. Nevertheless, we have notion that if VFC and the clock gating are used together in an on-chip single platform, then there can be possible reduction of PSN.

7. Conclusion

This chapter is dedicated for the discussion on what are the implications of on-chip PSN and the corresponding ways to mitigate and suppress it quantitatively. Over the years, the impact of PSN has been so adverse on the functioning of on-chip components, that the circuit designers and researchers had to narrow down the best possible options for PSN mitigation and classify these options as circuital and non-circuital approach. However, there is not much scope of deterministic non-circuital solution to the issues brought up by PSN because of the constant upgrading in the semiconductor manufacture.

In case of the circuital approach, it has been pointed out that by manipulating the on-chip clock activity, the major factors of PSN (i.e., i(t) and di/dt) can be in control. Until now, the introduction of VFC and clock gating individually inside the on-chip clock section has been the two most important circuital solutions for the PSN problems. However, this PSN controllability can be further enhanced by incorporating the VFC and clock gating together in place of the conventional system clock assuring that the chip functionalities are not compromised.

Acknowledgements

Conflict of interest

The authors declare no conflict of interest.

Author details

Pritam Bhattacharjee[1], Prerna Rana[2] and Alak Majumder[1*]

1 Integrated Circuits and Systems (i-CAS) Laboratory, National Institute of Technology (NIT), Arunachal Pradesh, Yupia, Arunachal Pradesh, India

2 VLSI Design Laboratory, National Institute of Technology (NIT), Manipur, Imphal, Manipur, India

*Address all correspondence to: alak@nitap.ac.in

IntechOpen

References

[1] Gray PR, Hurst P, Meyer RG, Lewis S. Analysis and Design of Analog Integrated Circuits. California, USA: Wiley; 2001

[2] Arabi K, Saleh R, Meng X. Power supply noise in SoCs: Metrics, management, and measurement. IEEE Design and Test of Computers. 2007;**24**(3):236-244

[3] Weste NH, Eshraghian K. Principles of CMOS VLSI design: A systems perspective. NASA STI/Recon Technical Report A. 1985;**85**

[4] Hook TB, Breitwisch M, Brown J, Cottrell P, Hoyniak D, Lam C, et al. Noise margin and leakage in ultra-low leakage SRAM cell design. IEEE Transactions on Electron Devices. 2002;**49**(8):1499-1501

[5] Nourani M, Radhakrishnan A. Power-supply noise in SoCs: ATPG, estimation and control. In: IEEE International Conference on Test. IEEE; 2005. p. 10

[6] Larsson P. Power supply noise in future IC's: a crystal ball reading. In: Proceedings of the IEEE 1999 Custom Integrated Circuits Conference (Cat. No. 99CH36327). IEEE; 1999. pp. 467-474

[7] Pant S, Blaauw D, Zolotov V, Sundareswaran S, Panda R, Panda R. Vectorless analysis of supply noise induced delay variation. In: Proceedings of the 2003 IEEE/ACM international conference on Computer-aided design. IEEE Computer Society; 2003. p. 184

[8] Tirumurti C, Kundu S, Sur-Kolay S, Chang YS. A modeling approach for addressing power supply switching noise related failures of integrated circuits. In: Proceedings of the conference on Design, automation and test in Europe-Volume 2. IEEE Computer Society; 2004. p. 21078

[9] Lin L, Jain S, Alioto M. Reconfigurable clock networks for wide voltage scaling. IEEE Journal of Solid-State Circuits. 2019;**54**(9):2622-2631

[10] Rabaey JM, Chandrakasan AP, Nikolic B. Digital Integrated Circuits. Englewood Cliffs: Prentice hall; 2002

[11] Mistry KR, Fox TF, Preston RP, Arora ND, Doyle BS, Nelsen DE. Circuit design guidelines for n-channel MOSFET hot carrier robustness. IEEE Transactions on Electron Devices. 1993;**40**(7):1284-1295

[12] Dharchoudhury A, Panda R, Blaauw D, Vaidyanathan R, Tutuianu B, Bearden D. Design and analysis of power distribution networks in PowerPC/sup TM/microprocessors. In: Proceedings 1998 Design and Automation Conference. 35th DAC. (Cat. No. 98CH36175). IEEE; 1998. pp. 738-743

[13] Popovich M, Mezhiba A, Friedman EG. Power Distribution Networks with on-Chip Decoupling Capacitors. Berlin, Germany: Springer Science & Business Media; 2007

[14] Gowan MK, Biro LL, Jackson DB. Power considerations in the design of the Alpha 21264 microprocessor. In: Proceedings of the 35th annual Design Automation Conference. ACM; 1998. pp. 726-731

[15] Becker WD, Eckhardt J, Frech RW, Katopis GA, Klink E, McAllister MF, et al. Modeling, simulation, and measurement of mid-frequency simultaneous switching noise in computer systems. IEEE Transactions on Components, Packaging, and Manufacturing Technology: Part B. 1998;**21**(2):157-163

[16] Sham KJ, Harjani R. I/O staggering for low-power jitter reduction. In: 38th European Microwave Conference. IEEE; 2008. pp. 1226-1229

[17] Musumeci S, Raciti A, Testa A, Galluzzo A, Melito M. A new adaptive driving technique for high current gate-controlled devices. In: Proceedings of 1994 IEEE Applied Power Electronics Conference and Exposition-ASPEC'94. IEEE; 1994. pp. 480-486

[18] Su DK, Loinaz MJ, Masui S, Wooley BA. Experimental results and modeling techniques for substrate noise in mixed-signal integrated circuits. IEICE Transactions on Electronics. 1993;**76**(5):760-770

[19] Bhooshan R. Power grid design in an integrated circuit. United States patent application US 11/163,520; 2007

[20] Vazgen SM, Karo HS, Avetisyan VA, Hakhverdyan AT. On-chip decoupling capacitor optimization technique. In: IEEE 37th International Conference on Electronics and Nanotechnology (ELNANO). IEEE; 2017. pp. 116-118

[21] Kirolos S, Massoud Y, Ismail Y. Power-supply-variation-aware timing analysis of synchronous systems. In: IEEE International Symposium on Circuits and Systems. IEEE; 2008. pp. 2418-2421

[22] Branson CN. Variable frequency microprocessor clock generator. United States patent US 4,819,164; 1989

[23] Fischer T, Desai J, Doyle B, Naffziger S, Patella B. A 90-nm variable frequency clock system for a power-managed itanium architecture processor. IEEE Journal of Solid-State Circuits. 2005;**41**(1):218-228

[24] Fan Q, Zhang G, Hu W. A synchronized variable frequency clock scheme in chip multiprocessors. In: IEEE International Symposium on Circuits and Systems. IEEE; 2008. pp. 3410-3413

[25] Tierno J, Rylyakov A, Friedman D, Chen A, Ciesla A, Diemoz T, et al. A DPLL-based per core variable frequency clock generator for an eight-core POWER7™ microprocessor. In: Symposium on VLSI Circuits. IEEE; 2010. pp. 85-86

[26] Huang Z, Luong HC. An 82-to-108GHz–181dB-FOM T ADPLL employing a DCO with split-transformer and dual-path switched-capacitor ladder and a clock-skew-sampling delta-sigma TDC. In: IEEE International Solid-State Circuits Conference-(ISSCC). IEEE; 2018. pp. 260-262

[27] Bhowmik S, Deb D, Pradhan SN, Bhattacharyya BK. Reduction of noise using continuously changing variable clock and clock gating for IC chips. IEEE Transactions on Components, Packaging and Manufacturing Technology. 2016;**6**(6):886-896

[28] Bhowmik S, Pradhan SN, Bhattacharyya BK. Power supply noise reduction of multicore CPU by staggering current and variable clock frequency. IEEE Transactions on Components, Packaging and Manufacturing Technology. 2018;**8**(5):875-882

[29] Majumder A, Bhattacharjee P. Current Profile Generated by Gating Logic Reduces Power Supply Noise of Integrated CPU Chip. In: IEEE International Symposium on Nanoelectronic and Information Systems (iNIS). IEEE; 2017. pp. 224-228

Section 5

Multimedia Communication

Chapter 11

A Survey on Adaptive Multimedia Streaming

Kunwar Pal, Mahesh Chandra Govil, Mushtaq Ahmed and Tanvi Chawla

Abstract

Internet was primarily designed for one to one applications like electronic mail, reliable file transfer etc. However, the technological growth in both hardware and software industry have written in unprecedented success story of the growth of Internet and have paved the paths of modern digital evolution. In today's world, the internet has become the way of life and has penetrated in its every domain. It is nearly impossible to list the applications which make use of internet in this era however, all these applications are data intensive and data may be textual, audio or visual requiring improved techniques to deal with these. Multimedia applications are one of them and have witnessed unprecedented growth in last few years. A predominance of that is by virtue of different video streaming applications in daily life like games, education, entertainment, security etc. Due to the huge demand of multimedia applications, heterogeneity of demands and limited resource availability there is a dire need of adaptive multimedia streaming. This chapter provides the detail discussion over different adaptive multimedia streaming mechanism over peer to peer network.

Keywords: peer-to-peer network, video-streaming, multimedia, adaptive, quality of services

1. Introduction

The drastic improvement in communication networks over the last few decades has led to a rapid increase in the demand for data transmission. A rise in the demand and supply of media content has been witnessed due to the variety of multimedia platforms available over the internet. Some of the platforms that provide multimedia content over the internet are YouTube [1], Net TV [2], and IPTV [3], etc. Among all data that is transmitted over P2P networks; over 90% of the data is due to video transmission. YouTube is the second most popular site according to the Alexa Ranking System [4]. Demand for multimedia content is increasing due to its widespread use in the communication, entertainment and education sectors [5]. Two different approaches used for video transmission over the network are Video on Demand (VoD) and live video streaming. VoD approach is more competent and convenient for the user, as the user can demand the video whenever they choose to rather than watching it at a particular time. No real time constraints are applied and no broadcast time is fixed only during which the video can be watched [6]. But live video streaming approach follows a specific broadcast time. If a user wants to watch

the video, they have to watch it during that broadcast time. The network require-ments for live video streaming are more as compared to that for video on demand. Some of the famous commercial P2P live video streaming systems are UUSee [7], Sopcast [8], PPLive [9]. In live video streaming, video transmission is more complex as compared to that in VoD. The complexity of the network has increased over time due to the increase in demand for live streaming. **Figure 1** shows the categorization of the video traffic over the Internet. Overall Compound Annual Growth Rate (CAGR) of traffic is 24% between the years 2016 and 2021. From the figure, it is also clear that demand for Internet video increased from 51 to 67.4% in the same time span.

Further, this Internet video was also subcategorized to understand the demand of different video categories over the Internet. From the **Figure 2**, it is easily observed that there is a huge demand for live Internet video and there is 10% incre-ment in it in the period of 2016–2021. Some of the major highlights of video over Internet is describe in (**Table 1**).

Streaming or media streaming is a technique for transferring data so that it can be processed as a steady and continuous stream. Video streaming allowed to play the video without completely downloading the file. The Distinguishing distribution techniques for multimedia distribution applies specifically to telecommunication networks, as most of the delivery systems are either inherently streaming (e.g., radio, television) or inherently non-streaming (e.g., books, video cassettes, audio CD's).

1.1 Classification of video streaming

Generally, classification of media over the Internet can be done into two ways Video on Demand (VoD) and Live streaming [10].

1.1.1 Video on demand

Video on demand is a recent exploitation in information storage and com-munication. Entertainment has undergone several positive changes in the last decade. Video on demand is one of them. It is a technology which combines user requirement, computer network, communication, and video engineering. Video on demand is a system which works like cable television. Contents are pre-recorded and stored at servers. The server may be at one physical location or distributed at different locations. Content is buffered or downloaded before actual play. Video on demand is not so time critical approach like live streaming. Therefore, video can be paused or resumed during the play. Every user is free to choose the program of their demand. It means programs are broadcast according
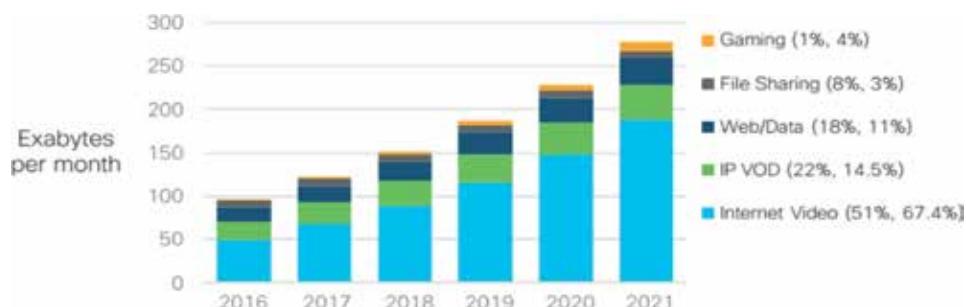
Figure 1.
Global IP traffic by application category [9].
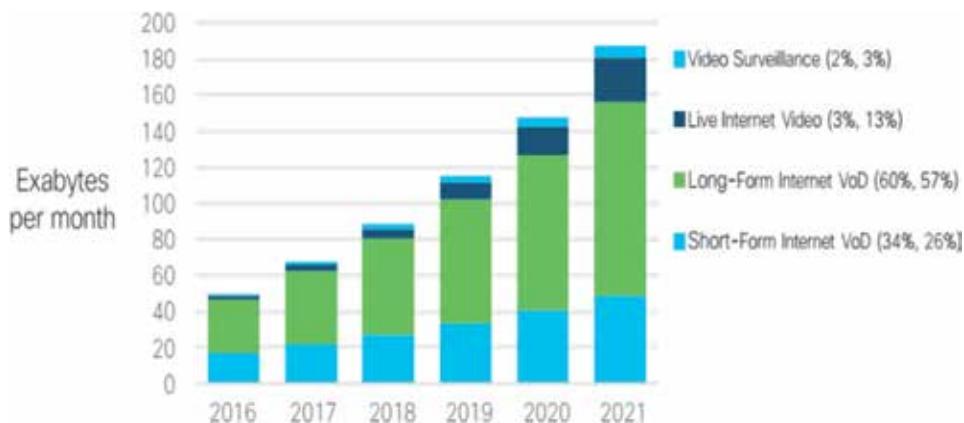
Figure 2.
Global internet video by sub-segment [9].

S. no	Property	Effect
1	Quantity of video	It will take approx. 5 million years to individual to watch video, Which will cross global IP in each month in 2021
2	Traffic of video	Video traffic will be approx. 73% up between 2016 and 2021.
3	Live video	Live video will grow 15-fold between 2016 and 2021.
4	Virtual reality and augmented reality traffic	Will increase 20-fold between 2016 and 2021
5	Internet video to TV	grew 50 percent in 2016
6	Consumer Video-on-Demand (VoD)	Traffic will nearly double by 2021
7	Content Delivery Network (CDN) traffic	It will carry 71 percent of all internet traffic by 2021

Table 1.
Video highlights.

to user requirements. A different user can ask for different services. The user can watch the programs according to their convenience. It may be weekdays or weekends at their choice timing. VoD also provides some special functionality. Users can forward or backward the program according to their convenience. In video on demand a client can download the video and may seek new play position according to his interest. Because the service is new it lacks universal standardization. **Figure 3** shows the simple example of video on demand. After capturing the live event from the camera encoding take place. Now encoded video is stored for further demand, and through the streaming server, video in delivered to requester peers A, B and C.

1.1.2 Live streaming

Live streaming is a video platform that allows users to view any video in real time. Broadcast video content using an audio/video communication media through the Internet. Live stream, can be described as the streaming platform that permits user to view and simulcast video content by the help of any capturing tools like

camera or audio recorder through the medium commonly known as the Internet. Live streaming can be described as the act of telecasting any live event which is being performed at real-time in any place by the medium of the Internet. It is not possible for each and every person to be present at many events going on simultaneously at different places. For this, live streaming plays a very important role in entertaining people by broadcasting the live event. For live streaming, a very common example is cricket matches being played at different places which are being telecasted for people entertainment.

Live video streaming is time critical as compared to video on demand. It's a delay sensitive service. The user can not choose the program at their convenience timing. The event is going on and it is broadcasted to all users. So users don't not have special features like fast forward or slow the live events. Packet sizes of video which are being sent from server to viewers are not predefined. It makes live streaming more complicated than video on demand. Transmission of the video is very crucial factor in case of live video streaming. Frames should be reached at the destination before its deadline. After deadline frames are useless. So the frames which are urgent should be reached first before other frames. Otherwise, there is an interruption in live video streaming. In video on demand, first of all video is stored and further distributed according to the demand of client. Due to the same, implementation complexity of the video on demand is less as compare to live streaming. **Figure 4** shows the simple example of live streaming. After capturing the live event from the camera encoding take place. Now the encoded video is transferred to a streaming server. The streaming server further forwards the video to requester peers A, B and C. A detail comparison of VoD and live streaming is given in **Table 2**.
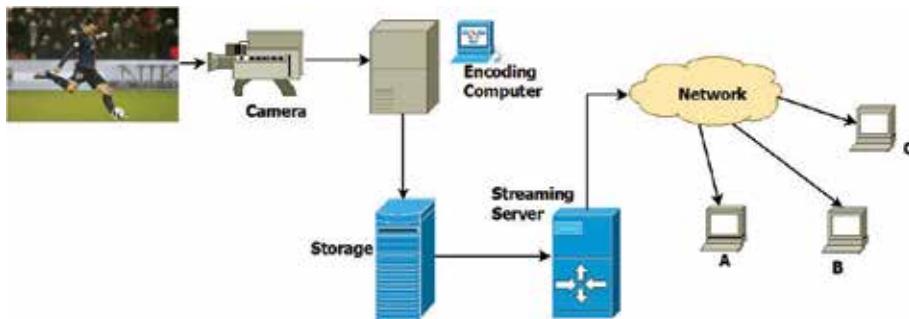
Figure 3.
Video on demand.
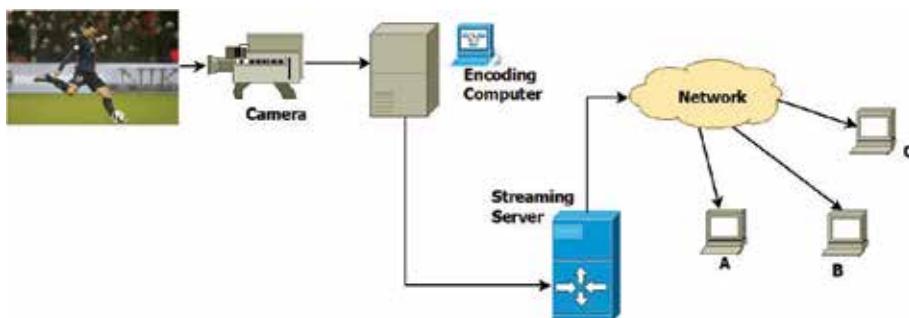
Figure 4.
Live streaming.

S. no	Live streaming	Video on demand
1	In a live streaming session, a live video content is disseminated to all users in real-time.	Video-on-demand users enjoy watching whatever video clips whenever they want
2	The video playback on all users is synchronized.	The playback of the same video clip on different users are not synchronized
3	All the users watching the same video and same segment.	If the users are watching the same video still the segment may be different
4	Less Flexible and convenience to users	VoD is more flexible and convenience to users
5	Time Complexity is more	Time Complexity is less
6	Space complexity is less	Space Complexity is more
7	Generally broadcast or multicast	Generally Multicast or Unicast
8	Economical	Required additional hardware and software

Table 2.
Comparison between live streaming and video on demand.

1.2 Architectures

The architecture use for any system plays very important role in its performance. Therefore, a good architecture is also required to efficiently support video streaming applications. The architectures used for such systems have to support broadly buffering and streaming of audio, video and related data. Client server architecture is conventional and most popular solution for both VoD and live streaming. However architecture is not appropriate due to the scalability and Quality of services of different parameters. This motivates the research community to search for better architecture and to most commonly used architectures are Content Distributed Network (CDN) and Peer to Peer (P2P) network discussed below.

1.2.1 Content distributed network (CDN) architecture

The CDN architecture is a distributed architecture in which data is distributed over the different server. Instead of storing the whole data at an original server, data is distributed at edge or surrogate server as shown in **Figure 5**. A hierarchy of server is created over the network [11]. When a client wants to communicate it will send a direct request to nearest edge server instead of an original server. If the server is free, it responds back according to data it has for the client. The client gets the response and according to response if it is positive, client start to see the live video. But if the response is negative client has to search another server and repeat the process again until it gets a positive response from the server. The server can share more than one program with multiple clients according to its resource capacity. To overcome the problem of a flash crowd in client-server architecture a new approach called cooperative networking (Coop Net) is discussed [5]. The drawback of Coop Net is that the client cannot work for long duration. Coop net uses distributed streaming and distributed coding approach for transmission of live video streaming. Using the approach robustness of the system is also increased. Control over-heard on the client side is minimum and easy to handle due to centralized behaviors of the server. Content distributed network provides reliability. In the client-server architecture, the resource utilization of the client is minimized. Scalability of the content distributed network is a major issue and increases the cost at the server side as the peer increase in the network. Due to the problem of scalability, cost and utilization of peer's resources researcher are attracted towards peer to peer network.

1.2.2 Peer-to-peer architecture

"Peer-to-peer (P2P) computing or networking is a distributed architecture that partitions tasks or assignments between peers. Peers are similarly advantaged, equipotent contributors in the application" [12]. The prime focus of P2P network is to fair distribution of the content without any extraordinary provision from the network. Aim of P2P network is to reduce the overhead from the server and uses the upload bandwidth of the users. Some of the formal definitions of P2P networks as defined in [13–15].

Figure 6 shows a tree overlay of peer to peer network with 13 peers and a streaming server. Peer 1 and 2 are receiving the content from the server directly while other peers receive it from their immediate parents instead of direct from server. So upload bandwidth of peers is also used to further upload the content in the network. A comparison of P2P and CDN is given in **Table 3**.

Working concept of P2P streaming system is same as BitTorrent [16], parent-child relation is stablishes between the peers from the set of neighbor peers. Landmark implementation of p2p with live video streaming is CoolStreaming [17], and the different issued relate to the implementation of CoolStreaming is disused
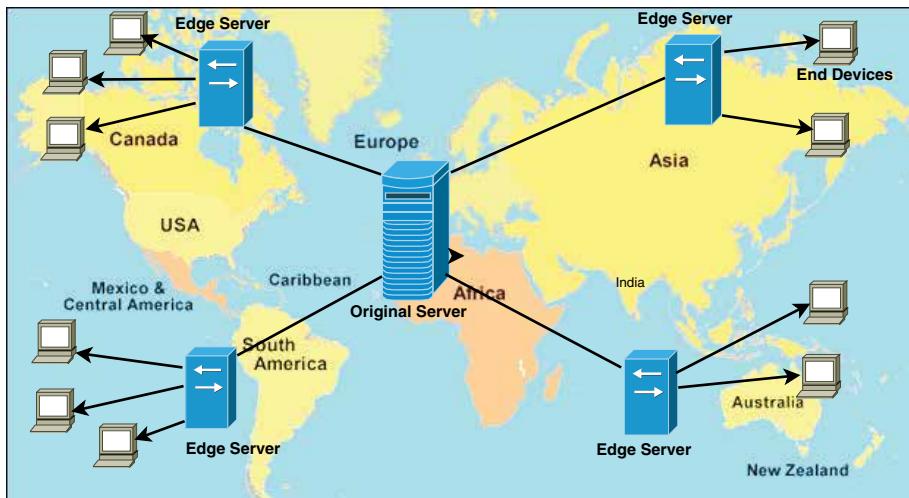
Figure 5.
Content distribution network.
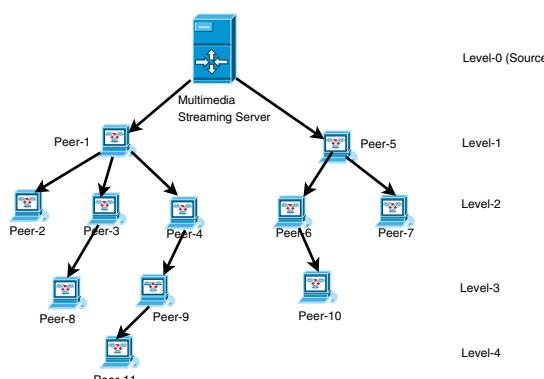
Figure 6.
Peer-to-peer network.

Property	P2P	CDN
Resource utilization	Maximized	Less
Peer role	Peer work as a both client and server	Specific role for client and server
Scalability	Easy and easily scalable	Limited
Cost	Low	High
Monitoring of content provider	Difficult	Easy
Stability	Low	High
Service capability	High	Limited
User management	Distributed	Centralized
QoS	Best-effort, Cannot be controlled	Can be guaranteed
Peer authentication	Distributed or no authentication	Centre node authentication
Services	Heterogeneous	Homogeneous throughout the network
Reliability	Low	High

Table 3.
Comparison between P2P and CDN.

by N.F. Huang et al. [10] and further visited in [18, 19]. PPLive [20], Sopcast [21], UUSee [22], Bittos [23], PPTV [24] and Mol's system [25] are some of the influential p2p streaming systems.

1.2.3 Hybrid architecture

In hybrid architecture both CDN and P2P architectures are combined to enhance the efficiency. A content distributed network is more reliable. Even if server is damaged peers can receive the media content from other servers. Peer to peer network is easily scalable and works better when the number of peer are more. Property of both the architectures are combined and a new hybrid architecture comes to light to leverages the characteristics of both the architectures [26]. In hybrid architecture, if a peer fails to deliver the requirements then reliable server takes over and provides the required media content to all the requester peers. Until the peer searches the new parent peer the quality maintenance in the network is the responsibility of the reliable server. Design issues of a hybrid system is discussed by Cui and Nahrstedt [27]. Servers are considered with large storage and bandwidth and peers are considered with limited storage and bandwidth. The link between the peers in the network follows the property of same as in P2Pand CDN. **Figure 7** gives an example of hybrid architecture, 15 peers are in the network. Peer 1-4 are in P2P network and receive the content from the server2. Server 1-6 is the part of the content distributed network.

Landmark implementation of live video streaming in P2P network is CoolStreaming [13]. Some other popular techniques for P2P are PPTV [14], Mol's system [15] and Bitos [16]. Different implementation issues regarding CoolStreaming have been discussed in [17]. The design issues of DoNet/CoolStreaming are briefly described [10]. For overlay construction in CoolStreaming swarm based approach is used. An empirical study of CoolStreaming [18] reveals that distribution of media content can also affect the performance of the P2P network.
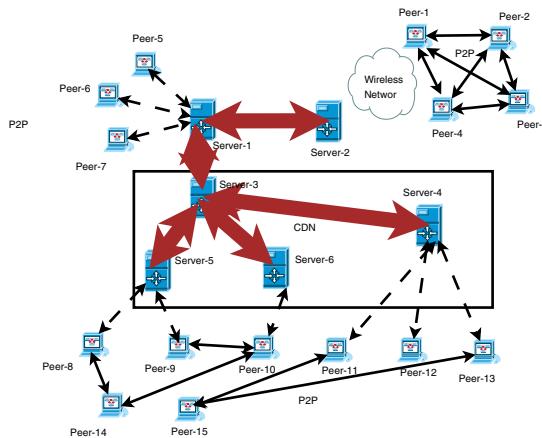
Figure 7.
Peer to peer and Content Delivery Network.

Some of the most challenging tasks in Multimedia streaming is selfish peer removal, flash crowd and data scheduling due to the frequently changed behavior of peers and network. This chapter primary focus on adaptive multimedia streaming. Further organization of the chapter is as follows. In Section 2, the work on adaptive multimedia streaming has been discussed. Section 2.1, gives the adaptive scheduling scheme, section 2.2 adaptive peer selection. Section 2.3 illustrates the various rate adaptation mechanisms. Last, the chapter concludes in Section 3.

2. Adaptive Multimedia streaming

Multimedia streaming architecture can be considered as a request response model and the demand of each client is not fixed also the network configuration can be changes dynamically. So static demand supply mechanism cannot provide the solution of multimedia streaming problem. Multimedia streaming model should be adaptive in nature. Different adaptation techniques can be applicable and some on them is discussed below in detail.

2.1 Adaptive Scheduling

The scheduling strategy changes with any change in the network configuration. So, peer selects the different block of chunk according to its current network configuration or current peer demand. In this section, we discuss the different adaptive parameters which can affect the scheduling strategy and the different adaptive scheduling schemes. In **Figure 8**; different adaption parameters are discussed. Passive adaptation parameters are those which are pre-defined and do not change in the network for a specific peer. While active adaption parameters are the ones which change as the network condition changes. For a peer, active adaption parameters are dynamic in nature while, the passive adaption parameters are static in nature. This section explains how according to adaption, the choice of scheduling scheme for different parameters changes. Active adaption is dynamic so it is crucial for scheduling scheme and different parameters like peer, congestion etc. affect the scheduling.

The relationship between sender and receiver may worsen due to improper resource adjustment. Lack of network knowledge also affects the network performance. Zheng et al. [28] described an adaptive approach, which uses push based scheduling scheme to solve the problem of the unstable relationship between
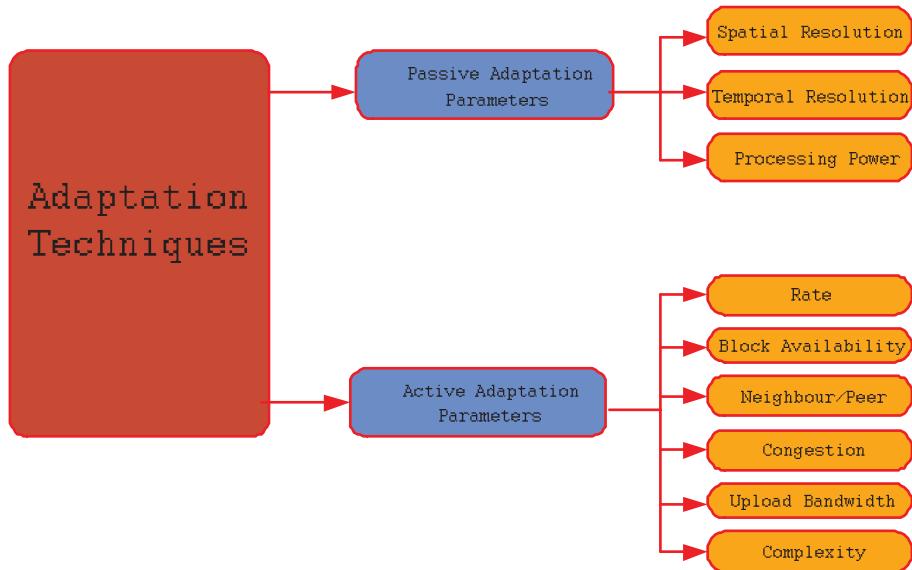
Figure 8.
Adaptation Techniques.

demand and supply [28]. This approach depends on forecasting. The system forecasts the utilization of resources i.e. which resources will probably be insufficient so that resources can be demanded early and scheduling becomes smooth. An Adaptive method is used to select the push peers and the push resources. Peers are chosen according to network conditions and among them, the best peers are selected. Heterogeneity of the network and end devices is a major factor in live video streaming. The different issues that arise are resolution and screen size, delay, variable download bandwidth and processing capability. According to user convenience, the author tries to improve video streaming for end users. This approach provides quality adaptive streaming using scalable video coding (SVC). SVC is a layer coding technique where the video is distributed at different layers i.e. one base layer and multiple enhancement layers [29]. Layer selection is a two-step mechanism, and Quality adaption is provided using layer selection. The first step is Initial Quality Adaption (IQA) and the second step is Progressive Quality Adaption (PQA). **Figure 9** depicts the Quality adaptive streaming architecture.

Initial Quality Adaption (IQA): As the name of the module suggests it is an initial decision for quality adaption. Parameters which influence the decision in IQA are screen resolution, available bandwidth and processing power. The initial quality set is assumed as (QS0). The value of QS0 is defined in Eq. (1), where d, t, and q are the total number of layers in spatial, temporal, and SNR dimensions.

$$QS0 \; \Delta = \big\{ (d, t, q) : \forall d = 0...D, t = 0...T, q = 0...Q) \big\} \qquad (1)$$

Progressive Quality Adaption (PQA): This module is executed periodically to improve the video quality according to network changes. It takes two real-time factors into consideration i.e. throughput and block availability. It also considers processing power to maintain peer resources. On the basis of above parameters, it calculates the net status adaption, bitrates adaption and complexity adaption. Lastly, it concludes with the final adaptive output [30].

The dynamic requirement of peers is satisfied by parameters which are vital for the end user. Approach provides results considering the dynamic effect of nature,
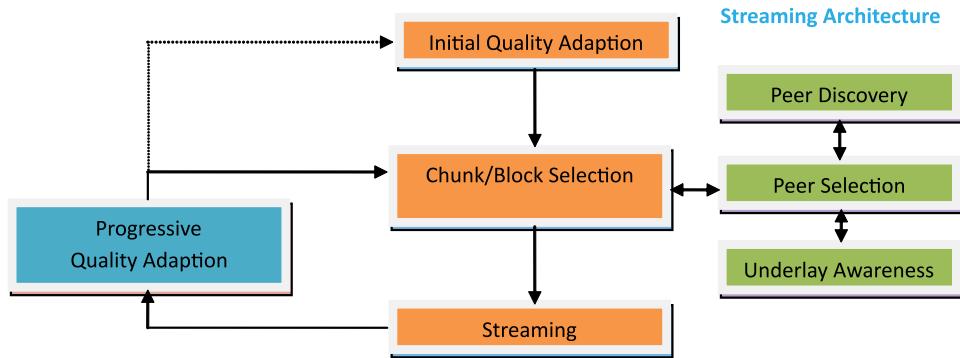
Figure 9.
The quality adaptive P2P Streaming Architecture.

and thus it provides better adaptive results regarding block availability and response time for the reply. A solution to the scheduling problem is to find the sender peer and the resource. Many different paths maybe available between the source and destination in the case of mesh overlay. Thus, finding the best path between the sender and receiver also plays a major role in P2P networks. The approaches to path selection using genetic algorithm are used to take care of the same [31]. These methods use mutation and crossover procedures with series of the population. Peers find information about the physical path and this information works as a fitness function for decision making. Traceroute command is used for calculating physical path information; this command dynamically finds changes in the network. Approach adaptively provides the best available path between the sender and the receiver. Using such path selection approach has a positive effect on the network performance regarding the end-to-end delay. **Figure 8** shows a classification of the adaptation techniques into two parameters on which they are based. Out of these different parameters adaptive peer selection and adaptive rate selection have a major impact on P2P network, so these two adaptation mechanisms are discussed in detail.

For providing the full bandwidth utilization of the peers available in the specific region E. Karayer provides a multi-generation packet scheduling approach [31]. Different classes according to the desired quality of services are created. Further two scheduling approaches are developed in such a way that first approach consider primarily on single generated problem, while second approach focuses on dynamic programming. Solution of the approaches are achieved using pseudo-polynomial-time. Improved bandwidth efficiency is achieved and verified thorough the simulation results [32]. Another scheme which focuses on multi-chunk and peer selection is defined in [33]. Focus of the approach is to provide usefulness of peers and chunks according to the playback probability of each chunk and streaming rate. Available bandwidth of each peer in the network also play a crucial role for scheduling scheme. Considering this factor J. Zhang [34] provides new approach for the selection of scheduling scheme. Real-time congestion of the environment is also considered for the selection of scheduling. Through the congestion peers are considered as busy, sense busy and idle. As a comparative analysis with CoolStreaming approach provides better Quality of Services (QoS) with respect to delay and throughput. Another approach defined in in [35] uses the 324 bandwidth information of every available peer in the network for selecting a 325 scheduling scheme.

2.2 Adaptive Peer Selection

The P2P network approach is scalable, so it can effortlessly be used in a video streaming network. But how a peer finds a perfect video chunk from its neighbor

peers is a challenging task in the P2P network. The author X. Liao et al. [36] focuses on this issue and tries to formulate a hybrid scheduling scheme for data transmission in P2P network. Some of the factors which affect scheduling scheme are:

1. Which part of data to schedule?

2. Who should be the parent for a particular data packet?

The OCTOPUS approach takes the second factor into prime consideration and uses a score method for parent selection between the available neighbors. The score values calculated between the peers maybe positive or negative and are thus categorized as positive score or negative score. If a client needs data from its neighbor's buffer cache, it is called a positive score. But if a neighbor requires data from its client's buffer cache it is called a negative score. In simple terms, data required by the client is positive score and data given by client is the negative score. Eq. (2) represents how a score is calculated, Eq. (2) Eq. (3), Eq. (4) and Eq. (5) are used in this process.

$$\text{Score}(k, j) = \text{Score}_{j \to k} + \text{Score}_{k \to j} \tag{2}$$

where

$$\text{Score}_{j \to k} = \sum_{i = \text{PMAX}_p}^{\text{PWANT}_k} S[i - \text{PMIN}_j] * \overline{S}[i - \text{PMIN}_k] \tag{3}$$

$$\text{Score}_{k \to j} = \sum_{i = \text{PMAX}_j}^{\text{PWANT}_j} S[i - \text{PMIN}_k] * \overline{S}[i - \text{PMIN}_j] \tag{4}$$

$$S[i] = \begin{cases} 0 & \text{empty} \\ 1 & \text{full} \\ 1 & \text{Scheduled} \end{cases} \tag{5}$$

The peers exchange buffer map among themselves to update their status. A Buffer map contains of 4 tuples <Pmin, Pmax, Pwant, S[i]>. According to the score value, each peer selects its parents. If score value of j is more for peer i, then peer i selects j as a parent. Scheduling cycle is assumed as T, while R(x) and W(x) as upload bandwidth and initial scheduling bandwidths for peer x. Then the capacity is calculated as P(x) = W(x)*T. Initial scheduling bandwidth of each peer is calculated as W(x) = Sbps/N, where Sbps is streaming rate for the video file and N are the total number of parents. After making the comparison between W(x) and R(x), the bandwidth is adaptively increased or decreased according to the network conditions. This is done so that the upload bandwidth of the peer can be fully utilized. The implementation done by the author shows that OCTOPUS approach gets the desired results. The main consideration of this approach is choosing the best parent peer. Selection of data is also a major concern in the network, but this approach doesn't consider this issue. Other issues like end device heterogeneity and network configuration have also not been discussed in this approach [36].

Pre-caching of video can solve the issue of selecting an optimal scheduling scheme upto a certain level. Different pre-caching approaches for distributed and centralized settings are given in [37]. Max-weighted independent set (MWIS)

method is used for the centralized scheduling setting. The message passing approach is used to find the value of the max-weighted independent set. For distributed scheduling setting, FlashLinQ link scheduling algorithm is used. Pre-caching at single level is not sufficient and can be further improved and the same is discussed by Hui-Hsiang Kao [38]. Hui-Hsiang Kao distribute the caching into two levels and named as prefix and suffix caching. Low startup delay and fast parent searching is achieved using prefix caching while adjustment of cache size is provide by suffix caching. Overall mechanism make sure the reduction in probability of chunk loss.

Peer selection is a prime concern to minimize the start-up delay and to provide high quality video. Round trip time (RTT) and available bandwidth both play a fundamental role in the P2P network. A combination of both parameters has been discussed in Adaptive and Efficient Peer Selection (AEPS) [39]. This approach combines both the factors i.e. RTT and Bandwidth, for video streaming. Different stages used in the architecture are Demand, Query & Rank and Verify. Demand stage is the core step in any P2P network. In this stage, a peer asks for a list of other available peers in the network who can play the role of parent for the client peers. Index server or tracer contains the list of existing peers in the network that have the same video. If there are too many peers in the network, instead of the full list server sends a partial list, and the rest of the list is calculated randomly. On receiving the list, the peer sends a query to all the available parents in the list. In response to the query, client receives a reply message. On the basis of this reply message, client finds the rank of all parent peers. APES scheme considers RTT as a principal factor; client uses this RTT value for ranking peers. Among all peers, if the RTT value of a peer is the highest then ranking of that peer is the lowest and based on the bandwidth further improvement must be made. In stage 3, the path defined by step 2 is verified according to the available bandwidth. The Parents peers are authenticated from the list according to their ranks. Through the experimental results, it is shown that AEPS is more efficient than the ABW based approaches and more accurate than the RTT based methods. APES performs much more efficiently in a general network [39].

The peers are very unpredictable in nature, and a peer cannot trust other peers for good data transmission so, trust is a crucial factor for good data transmission. On the basis of trust model a novel approach, trustworthy-based efficient multipoint relays (TEMRP) is described [40]. Method is described for mobile P2P networks which are a major component of P2P network. Approach tries to improve communication overhead between peers. A comparison is done between the TEMRP, multipoint relay (MPR) and pure flooding methods which reveal some useful results. MPR and flooding approach are very basic strategies for mobile P2P networks. Using the cost model a comparison of the different methods is performed using the resource utilization and latency. Peer calculates trust values based on the past experiences and according to that, it finds the trustworthy peer. Only trustworthy peers are selected as the next relay peer for data transmission. Most trustworthy peer selects the second most trustworthy peer and thus forwards the data further. Broadcast peer also finds the leaf peer and assigns a trustworthy peer for them so data can reach towards the leaf peers. The format of the message is <p1# p1-1, p1-2... p1-m>, peer before #indicates that peer p1 sends data to other peers that are shown after #. Then different trust values are calculated, and cost model is used to derive values for the TEMRP. In the end the cost of relay peer, leaf peer, and the broadcast peer is calculated. The comparison of cost model with MPR model and simple flooding model reveals that energy cost and latency of TEMRP is better than the other two approaches i.e. the MPR and Flooding approach [41].

Divide and conquer approach can be used to find the appropriate peer dynamically. Network Utility Maximization (NUM) problem is formulated first and

second problems of NUM approach are solved by using Lyapunov drift plus penalty approach. The dynamic adaptive approach is used which decomposes the problem into two blocks. The first block is used to find the adaptive video quality by using helper selection mechanism. The second block is used for discovering the helper to peer rate. Thus this approach is adaptive in two ways: firstly for selecting of helper and secondly for rate adaption. Queue size also affects the performance of the network so in the approach the queues are assumed to be of the same length. For the purpose of buffering the pre-buffering and re-buffering approaches are used [42]. Networks in live video streaming have a demand for adaptive streaming. Every individual peer tries to improve the quality of media it receives. Peers try to receive the video as smooth as possible. HTTP Adaptive streaming (HAS) can be a solution for adaptive streaming [43] where a video is split into small segments, and then these segments are encoded at different rates. Differentiation between the rates is maintained according to the quality of the video. The decision of choosing any rate among the different possible rates is according to network conditions and the buffer status of the peer. If the network is less congested, then peer can request for better quality, but if the network is congested then, the peer has to adjust with low-quality video. The same approach is used if the buffer is full and if a peer can wait for better quality it will request for a better quality video. But if the buffer is empty and there is a situation of starvation, then the peer can only ask for low-quality video. The approach uses an optimization agent which monitors the network conditions and according to that it calculates the required optimization value and provides the best deal for it. Packet-based traffic estimation model is used for calculating traffic between the peers. A graph is created, and traffic of each path (edge) is calculated. The historical approach is used for calculating or estimating traffic between the peers. Then QoE-driven quality adaption technique is used for quality updating. This approach considers issues like rate and traffic but parameters like chunk utilization and chunk deadline which can also affect the network performance are not considered.

2.3 Adaptive Rate Selection

Only adaptive peer selection is not sufficient for improvement of a P2P network hence, overlay adaption and rate adaption are combined in a new scheduling scheme as described by Hao Luan [44]. On the basis of local information; rate controller applies rate adaption in this scheme. This approach is efficient in P2P networks for overlay construction and as a scheduling scheme, but it doesn't contemplate upon properties like heterogeneity and parameters like delay in the network. A fixed channel uses fixed rate for data transmission over P2P network. If the data rate is high, the chance of data drop increases due to lack of available bandwidth between the peers. But if the data rate is low, then there is a chance of video smoothing and video quality. So the data rate directly affects the performance of video streaming in P2P network. To solve the above problem, an approach for the self-adaptive multi-rate is discussed [45]. The architecture of the approach (MoSee) comprises of three main components: Tracker (register server), broadcasting server (Broadcaster) and the client (Consumer of data). The client maybe wired, or wireless, broadcaster has the responsibility to transfer data to the client according to the client requirements. It also registers the channel it created for the client, with the tracker. Client logs in the register server and finds the list of programs. After choosing the program, the client selects the channel and joins it. In MoSee, a stream of the channel is distributed into sub-streams, and each sub-stream is identified by a unique index. According to network condition and status of its buffer, the client can choose any sub-stream. The sub-stream can also be adaptively changed according to

variations in the network. In the starting, the network is unstable, so peer chooses low bit rate stream but after some time when the network becomes stable, the peer chooses the high data rate sub-stream. The concept of multiple sub-streams is introduced to decrease the degree of data sharing between peers who have different data sub-streams. For increasing the level of data sharing between such peers, a different approach is used. In that approach, peers that can download the high data rate also download the low data rate stream for poor peers. When the poor peer (the one with a low download bandwidth) needs the low rate sub-stream, then that peer can provide it the sub-stream. For selecting the neighbor peer, Round Trip time (RTT) value is used, and weights are assigned to different peers, after which the final selection is done.

Different factors like bandwidth heterogeneity, high transmission delay and churn of peers can be very tricky in a P2P network. A different prospective approach, server aided adaptive live video streaming that solves all these issues has been discussed by [46]. Multi-tree overlay approach is used for creating the overlay between the peers in P2P network. The video is sent by multiple peers according to the network conditions and available bandwidth. According to network state, streaming server switches bit stream between the multiple available streams. Rate shaping scheme is also applied to lower the rate for transmitting the stream to remove the effect of network congestion. Two stages are used for rate shaping in a distributed multi-tree overlay. The first stage is the calculation of allocated bandwidth, to compute this; a list of packets received and transmitted is created. According to remaining bandwidth, the most appropriate packet from the list is sent. In the second stage, the optimized scheduler is applied. On the basis of the impact on their performance, the packets are differentiated. This approach is suitable for application on the server side, but in P2P network the approach is not only at server side but also, the actual demand-supply exists between the peers. Another approach for controlling playback rate at run time is described by Maria [47]. This playback rate control method has been discussed in two different cases i.e. nominal and robust case. Remaining upload bandwidth of peers in the network is used for calculating the rate. It dynamically checks the remaining upload bandwidth of peers that are currently active in the network. A theoretical model is described, and the solution for playback rate control is provided.

The Higher rate is required to fulfill the demand of the receiver, and a queuing model can help with the same. If a proper queue is maintained and the proper input is provided to the queue, then appropriate rate can be received by the peers. For adaptive scheduling different queuing models are discussed in the approach like peer and server side [48]. In peer side queue, received data is classified into two classes F (Forwarding) and NF (Non-Forwarding). The forwarding class contains the data which should be forwarded to other peers, and non-forwarding class includes the data which belongs to peer. Thus, there is no need to forward it to other peers. At the server side, two queues are maintained one for data and the other for signal. Pull approach is used for retrieving data by the peers and optimization is done according to the queue status. A distortion rate approach has been discussed by [49], the approach is receiver driven on multipath streaming. First, the parent selection takes place, the most appropriate receiver driven approach is used to find the appropriate parent nodes. Parent selection takes care of optimal end-to-end delay path selection; then the stream flow equations are generated, and a stream flow graph is created. This solution of stream flow graph provides the best rate to the desired peer. To solve the problem of long start-up delay Z defined a low time delay scheduling scheme [50]. The three main methods that are used by the approach are described as copying, grouping based on parts and priority scheduling. The peer is distributed into different groups, and the peers within the same

group can exchange data with each other. A peer can also join another group after downloading the content from the first group. This approach is easy to implement as it also reduces the start-up delay in the P2P network.

3. Conclusion

Because of the properties like user-friendly nature and scalability the demand of video streaming is growing rapidly in the existing age. Demand of client and configuration of network is not static and that's force to moving towards the adaptation of streaming. Adaptive multimedia mechanism can provide the solution of scalability, network fluctuation, bandwidth availability and improvement in quality of services. So in this chapter we focus on different adaptive multimedia streaming mechanism. There are various mechanisms to provide it. Adaptive scheduling scheme is one of the crucial factors. Other factor which help in adaptive multimedia streaming are rate adaptation and peer adaptation. These factors directly impact the performance of video streaming and improve the overall quality of services for both sender and receiver as well as make the overall experience pleasant for client. Huge amount of work is already discussed but still there is a demand of appropriate adaptive scheduling scheme and the area is still open for further research.

Author details

Kunwar Pal[1]*, Mahesh Chandra Govil[2], Mushtaq Ahmed[2] and Tanvi Chawla[2]

1 Department of Computer Science and Engineering, National Institute of Technology Sikkim, India

2 Department of Computer Science and Engineering, Malaviya National Institute of Technology Jaipur, India

*Address all correspondence to: kunwar@nitsikkim.ac.in

IntechOpen

References

[1] Venkataraman V, Francis P, Calandrino J. Chunkyspread: Multi-tree unstructured peer-to-peer multicast. In: Proceedings 14th IEEE International Conference on Communication and Network Protocol—ICNP'06. 2006. pp. 2-11

[2] Jannotti J, Gifford DK, Johnson KL, Kaashoek MF, O'Toole JW Jr. Overcast: Reliable multicasting with on overlay network. In: OSDI'00 Proceedings of the 4th conference on Symposium on Operating System Design & Implementation. 2000. p. 14

[3] Awiphan S, Su Z, Katto J. ToMo: A two-layer mesh/tree structure for live streaming in P2P overlay network. In: 2010 7th IEEE Consumer Communications and Networking Conference, CCNC. Vol. 2010. 2010. pp. 1-5

[4] Index VN, Vni C, Vni C, Highlights VNIF. Cisco Visual Networking Index: Forecast and; 2017. pp. 2016-2021

[5] "YouTube." [Online]. Available: https://www.youtube.com/ [Accessed: 11-12-2018]

[6] "IPTV." [Online]. Available: https://www.iptvonline.ca/ [Accessed: 15 December 2018]

[7] "NetTv." [Online]. Available: http://nettv.com.np/nettv/ [Accessed: 25 November 2018]

[8] "Alexa." [Online]. Available: http://www.alexa.com/topsites [Accessed: 10 December 2018]

[9] Visual C, Index N, Vni C, Vni C. The Zettabyte Era: Trends and Analysis; 2017. pp. 2016-2021

[10] Huang NF, Chu YM, Chen YR. Design of a P2P live multimedia streaming system with hybrid push

and pull mechanisms. In: 2010 WRI International Conference on Communications and Mobile Computing, CMC 2010. Vol. 1. 2010. pp. 541-545

[11] "Akamai," 2017. [Online]. Available: https://www.akamai.com/ [Accessed: 15 December 2018]

[12] Li BLB, Yin HYH. Peer-to-peer live video streaming on the internet: Issues, existing approaches, and challenges [Peer-to-Peer Multimedia Streaming]. IEEE Communications Magazine. 2007;**45**(6):94-99

[13] Androutsellis-Theotokis S, Spinellis D. A survey of peer-to-peer content distribution technologies. ACM Computing Surveys. 2004;**36**(4):335-371

[14] Schollmeier R, Networks CT, Universität T. A Definition of Peer-to-Peer Networking for the Classification of Peer-to- Peer Architectures and Applications; 2002. pp. 2-3

[15] Camarillo G. Peer-to-Peer (P2P) architecture: Definition, Taxonomies, Examples, and Applicability. 2004;**5694**:1-26, Req. Comments

[16] Cohen B. Incentives build robustness in BitTorrent. Workshop on Economics of Peer-to-Peer Systems. 2003;**6**:68-72

[17] Xie S, Li B, Member S, Keung GY, Zhang X, Member S. Coolstreaming: Design, theory, and practice. IEEE Transactions on Multimedia. 2007;**9**(8):1661-1671

[18] Zhang X, Liu J, Li B, Yum TSP. CoolStreaming/DONet: A data-driven overlay network for efficient live media streaming. IEEE Conference

on Computer Communications. 2005;**3**:13-17

[19] Li B et al. An empirical study of the coolstreaming plus system. IEEE Journal on Selected Areas in Communications. 2007;**25**(9):1627-1639

[20] http://www.pplive.com, "PPLive," 2018. [Online]. Available: http://www.pplive.com [Accessed: 01 October 2018]

[21] http://www.sopcast.org, "SopCast," 2018. [Online]. Available: http://www.sopcast.org [Accessed: 15 October 2018]

[22] http://www.uusee.com, "UUSee," 2016

[23] Vlavianos A, Iliofotou M, Faloutsos M. BiToS: Enhancing BitTorrent for supporting streaming applications. In: INFOCOM 2006. 25th IEEE International Conference on Computer Communications. Proceedings. 2006. pp. 1-6

[24] PPTV, "PPTV," http://www.pptv.com/, 2016

[25] Mol JJD, Bakker A, Pouwelse JA, Epema DHJ, Sips HJ. The design and deployment of a BitTorrent live video streaming solution. In: ISM 2009 - 11th IEEE International Symposium on Multimedia. 2009. pp. 342-349

[26] Yang B, Garcia-Molina H. Comparing hybrid peer-to-peer systems. Proceedings of the 27th Intl. Conf. on Very Large Data Bases. 2001

[27] Cui Y, Nahrstedt K. Layered peer-to-peer streaming. In: Proceedings of the 13th International Workshop on Network and Operating Systems Support for Digital Audio and Video. ACM; 2003

[28] Zheng Y, Lin F, Yang Y, Gan T. Adaptive resource scheduling mechanism in P2P file sharing system. Peer-to-Peer Networking and Applications. 2015;**9**(6):2016

[29] Schwarz H, Marpe D, Wiegand T. Overview of the scalable video coding extension of the H.264/AVC standard. IEEE Transactions on Circuits and Systems for Video Technology. 2007;**17**(9):1103-1120

[30] Abboud O, Pussep K. Quality adaptive peer-to-peer streaming using scalable video coding. In: IFIP/IEEE International Conference on Management of Multimedia and Mobile Networks and Services (MMNS). Berlin Heidelberg: Springer; 2009. pp. 41-54

[31] Karayer E, Sayit M. A path selection approach with genetic algorithm for P2P video streaming systems. Multimedia Tools and Applications. 2015;**75**(23):16039-16057

[32] Hao P. Adaptive Packet Scheduling for Scalable Video Streaming with Network Coding; 2016. pp. 0-5

[33] Huang S, Izquierdo E, Hao P. Bandwidth-efficient packet scheduling for live streaming with network coding. IEEE Transactions on Multimedia. 2016;**18**(4):752-763

[34] Zhang J, Yang C, Zhang X. A high-bandwidth live streaming model in mesh-based peer-to-peer networks. IEEE Communications Letters. 2016;**20**(12):2390-2393

[35] Triningsih E, Bandung Y. Bandwidth allocation-aware scheduling algorithm for video-on-demand application over digital learning network. In: In ICT For Smart Society (ICISS), International Conference on, 2016. pp. 6-11

[36] Liao X, Jin H. OCTOPUS: A hybrid scheduling strategy for P2P VoD services. In: Proceedings of the 6th International Conference on Grid and Cooperative Computing, GCC 2007. 2007. pp. 26-33

[37] Kim J, Caire G, Molisch AF. Quality-aware streaming and scheduling for

Device-to-Device video delivery. IEEE/ ACM Transactions on Networking. 2016;**24**(4):2319-2331

[38] Kao H, Lee C, Kao Y, Wu P. I2CC : Interleaving two-level cache with network coding in peer-to-peer VoD system. Journal of Network and Computer Applications;**60**:180-191

[39] Hsiao TH, Hsu MH, Ben Miao Y. Adaptive and efficient peer selection in peer-to-peer streaming networks. In Parallel and Distributed Systems (ICPADS). In: IEEE 17th International Conference on Parallel and Distributed Systems (ICPADS). 2011. pp. 753-758

[40] Waluyo AB, Taniar D, Rahayu W, Aikebaier A, Takizawa M, Srinivasan B. Trustworthy-based efficient data broadcast model for P2P interaction in resource-constrained wireless environments. Journal of Computer and System Sciences. 2012;**78**(6):1716-1736

[41] Waluyo AB, Taniar D, Srinivasan B, Rahayu W, Takizawa M. Adaptive and efficient data dissemination in mobile P2P environments. In: Proceedings - 25th IEEE International Conference on Advanced Information Networking and Applications Workshops. Vol. 3. WAINA; 2011. pp. 861-866

[42] Bethanabhotla D, Caire G, Neely M. Adaptive video streaming for wireless networks with multiple users and helpers. IEEE Transactions on Communications. 2015;**63**(1):268-285

[43] Bouten N, Schmidt RDO, Famaey J, Latr S, Pras A, De Turck F. QoE-driven in-network optimization for adaptive video streaming based on packet sampling measurements. Computer Networks. 2015;**81**:96-115

[44] Luan H, Kwong KW, Hei X, Tsang DHK. Adaptive topology formation for peer-to-peer video streaming.

Peer-to-Peer Networking and Applications. 2010;**3**(3):186-207

[45] Yu L, Jin H, Jiang W, Liao G, Liao X. Self-adaptive schedule mechanism for peer-to-peer multi-rate live streaming system. In: Proceedings of the 14th IEEE International Conference on High Performance Computing and Communications. HPCC-2012—9th IEEE

[46] Zhang JF, Niu JW, Wang RG, Dong Y, La Wang H. Server-aided adaptive live video streaming over P2P networks. Journal of Signal Processing Systems. 2010;**59**(3):335-345

[47] Efthymiopoulou M, Efthymiopoulos N, Christakidis A, Athanasopoulos N, Denazis S, Koufopavlou O. Scalable playback rate control in P2P live streaming systems. Peer-to-Peer Networking and Applications. 2016;**9**(6):1162-1176

[48] Guo Y, Liang C, Liu Y. AQCS: Adaptive queue-based chunk scheduling for P2P live streaming. In: International Conference on Research in Networking. Berlin Heidelberg: Springer; 2008. pp. 433-444

[49] Liu SC, Chen TH. Enabling adaptive live streaming in P2P multipath networks. The Journal of Supercomputing. 2010;**53**(1):26-44

[50] Xu Z, Guan S. A low-time-delay scheduling algorithm for video on demand service over P2P networks. In: 2014 International Conference on Information and Communications Technologies (ICT 2014). 2014

Section 6

Network Routing Protocols

Chapter 12

Secure and Energy-Efficient Communication in IoT/CPS

Saad Alharthi, Princy Johnson and Martin Randles

Abstract

Secure and energy efficient routing protocol is fairly an open research despite a plethora of routing protocols has been proposed in the literature. However, most routing protocols specifically designed for resource constrained wireless devices, if not all, follow from the same perspective and almost have reached the maximum improvements. This chapter describes the design of cross-layer secure multi-hop zone routing protocol (MZRP) and a hybrid energy-efficient medium access control (MAC) featuring the benefits from both carrier sense multiple access (CSMA) and time-division multiple access (TDMA). MZRP employs the intelligent artificial neural network (ANN) self-organizing map (SOM) algorithm, which is performed at the coordinator or the base station (BS) to divide the area into multi-level zones. Then cluster heads (CHs) are chosen using k-medoids in each zone. The performance of MZRP is better in terms of energy efficiency compared to dual-hop and HT2HL as it extends the network lifetime using hybrid MAC and the security algorithm employed has less message update.

Keywords: multi-hop zone routing protocol (MZRP), self-organizing map (SOM), aggregation, internet of things (IoT), 6LoWPAN, artificial intelligence (AI), machine learning (ML)

1. Introduction

The architecture of the global network "Internet" will expand drastically to accommodate the connections of billion IP-enabled or non-IP smart wireless devices "things" at the network edge which brings up the era of the Internet of Things (IoT). Depending on the application, the data model can either be simple as small "things" are connected (e.g., wearables, smart home, etc.) or it can be complex as big "things" are connected (e.g., automotive, smart city, etc.). Hence, new architecture and technologies have been developed in the fifth generation (5G) to meet the sky-rocketing growth of the IoT [1].

The non-IP network edge utilizes an IoT architecture supported by IPv6 over low-power wireless personal area networks (6LoWPAN) to achieve seamless integration of different heterogeneous networks which enable the interaction between network edge "things" and the outside world "internet" [2]. A typical 6LoWPAN network consists of few hundreds or even thousands of low-cost and tiny intelligent sensor nodes (SNs) embedded with sensing transducers, microcontroller, and radio transceiver units, and all these units are powered by battery or other source of energy [3].

They form an ad hoc network among themselves and a gateway or 6LoWPAN edge router which acts as a centralized network coordinator or a base station (BS). The BS refers to gateway or 6LoWPAN edge router throughout this chapter. This type of network is known as wireless sensor network (WSN), which forms the major hardware devices of the IoT. Different WSNs can be seamlessly connected by routers to provide tunnels to the backbone [4].

The SNs in many applications have no fixed power source. Hence, they are powered by batteries where the replacement either be infeasible or might not be possible. The WSN's lifetime depends on the lifetime of batteries that operate the SNs. Thus, energy saving is the most important factor in designing WSNs, which mandates the necessity for energy efficient techniques to prolong the network lifetime of WSNs [5].

The WSN's lifetime maximization and the individual SN's energy conservation have been investigated by researchers, and several approaches have been proposed in the literature. The overall energy consumption in the SNs is due to sensing, processing, and communication. But radio transceiver has the highest energy expenditure, which involves transmit, receive, idle, and sleep modes. The energy consumption in sleep mode or low power listening (LPL) is very small; hence, the radio transceivers of SNs can be put in sleep mode, which can significantly increase the network lifetime as will be shown later.

This is governed by the performance of the protocols built across all layers of the open systems interconnection (OSI) protocol stack especially routing and medium access control (MAC) protocols within the SNs. Hence, routing and MAC protocols have critical impact on the network lifetime of WSNs. Therefore, careful design of energy efficient routing and MAC protocols is critical to conserve the wasteful energy and to utilize the useful energy wisely.

In a single-hop scenario, SNs can either send or receive messages to/from destination directly if they are within the transmission ranges of each other. If they are far away, i.e., the distance between a SN and a destination is more than the transmission range, and indirect communication using multi-hop is chosen where the message is relayed by several intermediate relays from a SN to a destination.

A simple approach is flooding where a SN sends a query packet for route discovery which propagates over the network until destination is reached. However, this results in broadcast storm due to the propagation of message even after route has been discovered. Thus, it is not suitable for energy constrained SNs as the energy consumption is hug. Data fusion employing passive clustering is another approach to achieve energy savings by processing data in network using machine learning and artificial intelligence (ML/AI) algorithms. It is an intuitive that the traffic and the energy consumption will be reduced significantly.

Hierarchical clustering has been proven to be an effective method for energy optimization in which SNs are organized into clusters. Each cluster has a cluster head (CH), which aggregates the data from cluster member nodes (MNs), and the CHs communicate directly to the BS using single-hop routing or relay the data through intermediate nodes using multi-hop routing. This combines the proactive approach within cluster (intra-cluster) and reactive approach across clusters (inter-cluster).

In battery-powered SNs, the design of low duty-cycle MAC protocol is crucial for energy savings. Most wireless technologies uses CSMA as channel access mechanism. Thus, the MAC layer of a SN is based on the IEEE 802.15.4 standard which defines the specifications of the lowest two layers of the OSI protocol stack; that is, the physical (PHY) and MAC layers. The PHY layer of the IoT protocol stack is based on the IEEE 802.15.4 standard; however, MAC layer is time-slotted channel hopping (TSCH) based on IEEE 802.15.4e standard to ensure high robust and reliable requirements in industrial applications [6]. Hence, these intelligent wireless

devices provide the backbone infrastructure for cyber-physical system (CPS) to monitor physical processes [1].

The distribution of SNs at different places and trying to connect them via wireless links make WSNs vulnerable to attacks. The types of attacks can be internal or external. External attacks can be prevented using link-layer encryption and authentication using global key (GK). However, internal attacks are difficult to defend since a compromised node inside the network is hard to detect as its identity can be verified using its private key (PK).

Hence, routing and forwarding packets can cause security problems as routing protocols are designed to exchange network topology between SNs to establish routes but not to provide defense against malicious attackers. Security in WSNs is still an open research as many questions remain unanswered. The existing network security approaches are either computationally complex, or resources are not efficiently utilized which makes the core security issue a big challenge under great resource constrained SNs.

The challenge is to design energy efficient routing and MAC protocols to prolong the network lifetime of WSNs and considers aspects related to security and data mining using artificial intelligence (AI) and machine learning (ML) algorithms under resource constraints. Thus, a careful design of routing and MAC protocols is required to enhance the data collection, improve security, and increase the lifetime of WSNs in thing-fog-cloud IoT architecture [6].

Our contribution is the design of a cross-layer secure multi-hop zone-based routing protocol (MZRP) and a hybrid energy-efficient MAC called X-SREM. MZRP is a hierarchical clustering routing and hybrid MAC which switches CSMA and TDMA based on data traffic. Low-duty cycle operation is achieved using cooperative strategies by shutting down the radio transceivers of SNs all times except during their assigned time slot. This novel design is useful to potentially achieve security and huge energy savings by eliminating the interference among densely deployed SNs and collision-free data transmission within clusters and reduces the traffic over the network by removing redundancy using data fusion algorithms.

The remainder of this chapter is organized as follows: Section 2 presents the related work. In Section 3, the problem is described. In Section 4, the models are defined. Section 5 describes our proposed cross-layer MZRP and hybrid MAC protocol. Section 6 presents the performance evaluation and analysis of the proposed protocol compared to two benchmark heterogeneous protocols dual-hop and HT2HL. The conclusion of this research is presented in Section 7.

2. Related work

Hierarchical clustering routing and low duty cycle MAC protocols have been proven methods to reduce the energy consumption of SNs and to eliminate the interference among them. Heinzelman et al. [7] proposed low energy adaptive clustering hierarchy (LEACH) and LEACH centralized (LEACH-C) [8], and they are the reference routing protocols in homogeneous hierarchical clustering category. LEACH has provided the basis for many later studies.

For instance, threshold sensitive energy efficient sensor network protocol (TEEN) and adaptive TEEN (APTEEN) are proposed in [9, 10], respectively. The election of CHs in these protocols is a random process. TEEN and APTEEN combine the hierarchical and data-centric approaches. Both reduce data transmission by performing data fusion at CHs, hence achieving large energy conservation. Hybrid energy efficient distributed clustering (HEED) was introduced in [11]. In HEED protocol, CHs are elected periodically based on the residual energy of the SNs and

the intra-cluster communication cost which incorporates cluster radius to adjust the transmission power for inter-cluster broadcast. The previous protocols and their variants are classified as homogeneous scheme protocols.

A stable election protocol (SEP) was proposed in [12] and distributed energy-efficient clustering (DEEC) proposed in [13] are the two well-known protocols for heterogeneous WSNs. In SEP and DEEC, the SNs are assumed to have different energy levels and processing power. Both assume m fraction of the total SNs are equipped with a factor α of additional resources compared to the normal SNs. Although both follow the same approach, DEEC considers the residual energy of SNs in the CHs selection process whereas SEP does not. SEP and DEEC and their variants are classified as heterogeneous scheme protocols.

Alharthi and Johnson [14] suggested HT2HL to deal with some drawbacks in LEACH. Abu Shiba et al. [15] proposed CH-LEACH which uses k-means algorithm to select the CHs, and Park et al. [16] proposed an efficient CH selection using k-means to maximize the energy efficiency of WSNs. Abdul Latiff et al. [17] compared the performance of optimization algorithms for clustering in WSN and validated that particle swarm optimization (PSO) achieves better network lifetime. The CHs are responsible to perform data fusion like typical artificial neural network (ANN) as we will discuss later.

The previous protocols and many of their variants [18] and a plethora other in the literature are single-hop hierarchical clustering routing for WSNs. This assumption is not realistic as the sink might not be reachable by all CHs or the network lifetime is drastically reduced as more energy is consumed in long distance. Multi-hop has more advantages and can solve this problem by choosing optimal routes in the tree formed between the leaves (CHs) and the root (BS).

Multi-hop LEACH finds paths using multi-hop approach like two-level LEACH [19] and MR-LEACH. In [20, 21], the authors proposed new routing techniques to minimize the next hop distance which reduces the energy. Al-Sodairi and Ouni [22] proposed EM-LEACH for load balancing and extended lifetime of WSNs. Modified RPL was proposed by Alharthi et al. in [23].

Routing protocols based on intelligent algorithms like ANN, fuzzy logic, PSO, genetic algorithm (GA), etc. have been surveyed in [24] for network lifetime optimization. A review given in **Table 1** is some existing single-hop and multi-hop and intelligent routing for WSNs.

Routing protocol	Clustering	No. of hops	Location aware	Scalability
LEACH	Distributed	Single-hop	No	Low
LEACH-C	Centralized	Single-hop	Yes	Medium
TL-LACH	Distributed	Two-hop	No	High
MR-LEACH	Distributed	Multi-hop	NO	Medium
HT2HL	Distributed	Two-hop	NO	Medium
CH-LEACH	Centralized	Single-hop	Yes	Medium
EM-LEACH	Distributed	Multi-hop	NO	Medium
BAR	Ant colony	Either-hop	Yes	Medium
GA-EECP	GA	Either-hop	Yes	Medium
FCH	Fuzzy logic	Either-hop	Yes	Medium
SIR	ANN	Either-hop	Yes	High

Table 1.
A summary of routing.

3. Problem statement

The design of energy efficient routing and MAC protocols aims to maximize the network lifetime (Tnet). The battery of a SN i will drain out under flow $\mathbf{f}$ = {fi,j} in a time (Ti). Hence, the objective is to minimize the total energy consumption Etot. The energy consumption in one cycle is the sum of energy in the active period (E_a) which depends on (e.g., transmit, receive or idle) and the sleep period (E_s). For n number of cycles, the $\mathbf{E}$tot = $\sum_{i=1}^{n} E_a + E_s$.

Most, if not all, previous protocols are based on LEACH and they inherited its drawbacks. Some drawbacks are highlighted in [14] and other important are discussed here. First, though there are enough CHs in some rounds, some rounds have larger, fewer or no CHs are elected. Second, the separation and compactness due to the randomness employed in the election of CHs. Finally, re-clustering consume large amount of energy but it is unnecessarily performed every round resulting in SNs with less remaining energy be elected as CHs after 1/p rounds where p is the percentage of CHs. This lowest energy rout will result in energy depletion of the nodes along the path and may lead to network partitioning [20].

The data aggregation is simple, which aims to reduce the number of messages. Hierarchical routing protocols play significant role to choose the SNs to perform data aggregation by constructing a loop-free spanning-tree. Hence, the latency decreases as the path length is decreased, and the accuracy and energy efficiency are increased. The resilience of aggregation can be improved using hierarchal techniques which provide defense against routing and aggregation attacks.

4. System models

The following section describes the network model for three-tier heterogeneous WSNs and defines the radio energy model which is used to compute the energy dissipated by the SNs.

4.1 Network model

Consider three-tier thing-fog-cloud IoT architecture as shown in **Figure 1** and WSN containing N nodes in tier-0 and fog nodes in tier-1. They are distributed uniformly at random in a 2D grid or a square sensor field (area A = LxL) to monitor the environment as shown in **Figure 1**.

There are two or more types of SNs might exist inside WSNs; i.e., full function devices (FFD), reduced function devices (RFD) and coordinator in Zigbee and Class A, B, and C in LoRA networks. Let S be the set of FFD SNs such that S = {S1, S2, ... , Si} where Si denotes the i-th SN, $1 \leq i \leq N$ represents the decimal ID of the SN. A SN can be equipped with one or more same or different types of sensors (e.g., temperature, pressure, etc.). The BS is static and located at the center with no limited energy. The SNs are unaware about their location and fixed after deployment and each has a unique ID as shown in **Figure 2(a)**.

The network in tier-0 can be denoted as a graph G = (V, E), where V is a set of vertices representing the SNs such that |V| = N and E $\subseteq$ V^2 is a set of edges that signify the bidirectional wireless links between a pair of SNs v_i and v_j. Hence, edges $(v_i, v_j) \in E$ if v_i and v_j are neighbors; i.e., the distance between them d (v_i, v_j) is less than the range of each other as shown in **Figure 2(b)**. Thus, E = {$(v_i, v_j) \in V^2$ | $v_i \neq v_j \wedge d (v_i, v_j) < R$}, where R is the communication range. We define $c_{i,j} (v_i, v_j)$ as the cost of energy on the edge between v_i and v_j assuming that the links are symmetric. That is, $c_{i,j} (v_i, v_j)$ = $c_{j,i} (v_j, v_i)$.
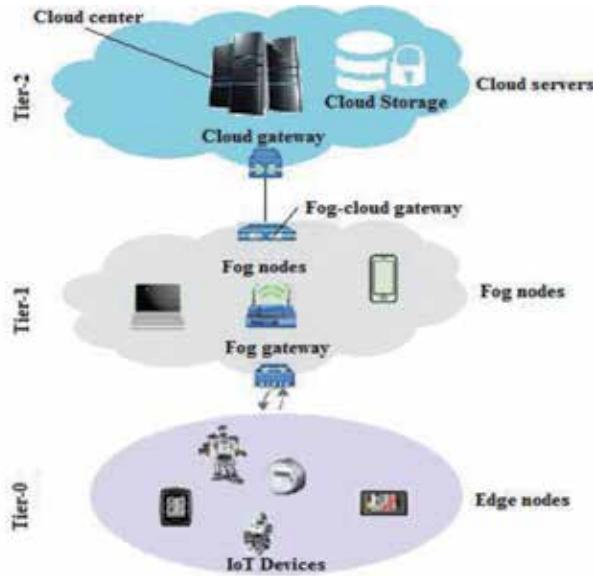
Figure 1.
The three-tier thing-fog-cloud IoT architecture.
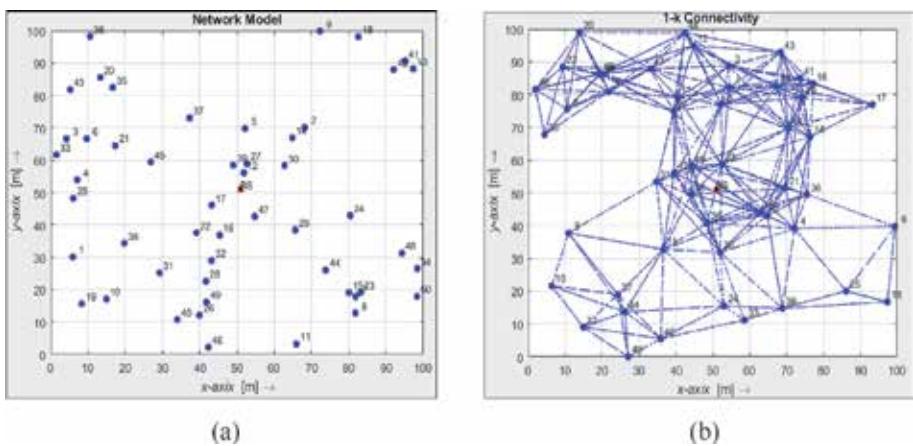
Figure 2.
Network model (a) deployment (b) connectivity.

Let G' = (C, E') be a subgraph, where C is a set of vertices representing the CHs such that |C| = k and E' $\subseteq$ C² is a set of edges that show the bidirectional wireless links between a pair of CHs C_i and C_j. Let $P(v_i, v_j) = (v_i, v_1, \dots, v_j)$ is the set of multiple paths in G' from C_i to C_j since other FFD SNs are in sleep mode and neither involve in routing nor act as fog nodes at this moment as shown in **Figure 3**.

4.2 Radio energy model

There are various models for radio energy consumption, though most of the researchers have used the energy model in [8]. The model for communication between SNs is depicted in **Figure 4**.

The energy consumption in the transmitter circuit E_{Tx} (B,d) to transmit B-bit message is given by
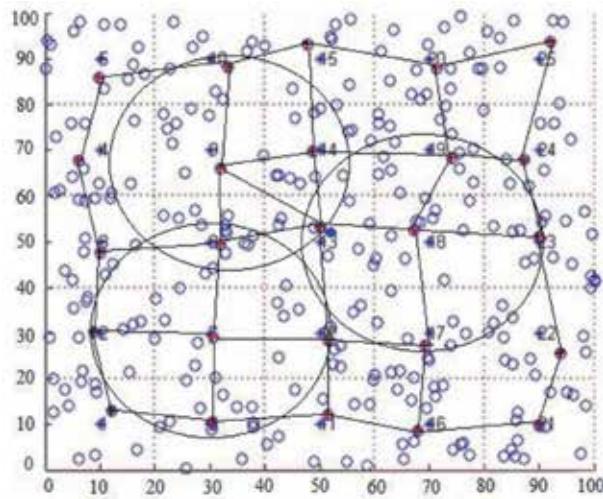
Figure 3.
Connectivity between clusters in multi-path.

$$E_{Tx}(B, d) = E_{elec} * B + \mathcal{E}_{amp} * B \qquad (1)$$

The E_{elec} depends on electronic circuits of the transceiver used in digital coding, modulation, demodulation, and other circuits. The $\mathcal{E}_{amp}$ is the energy dissipation in power amplifier, and it depends on power loss (PL) in the channel model used. The PL in free-space and log-distance models can be described using Friis equation as the ratio of the power transmitted to the power received (Pt/Pr) as given below:

$$\frac{P_t}{P_r} = \left[\frac{\sqrt{\lambda G}}{4\pi d} \right]^{\delta} \qquad (2)$$

where λ is the wavelength, G is the product of equal gains of transmit and receive antennas, d is the distance between antennas, and δ is the PL exponent which depends on the propagation environment. **Table 2** gives the typical values of δ.

Hence, d^2 is the power loss in free space ($\mathcal{E}_{fs}$) and d^4 is the power loss in two-ray or multipath fading ($\mathcal{E}_{amp}$).

$$\mathcal{E}_{amp} = \begin{cases} \mathcal{E}_{fs} * d^2, & d < d_o free - space \\ \mathcal{E}_{mp} * d^4, & d \geq d_o multi - path \end{cases} \qquad (3)$$

$$do = \sqrt{\frac{\mathcal{E}fs}{\mathcal{E}mp}} \qquad (4)$$

The values of $\mathcal{E}_{fs}$ = 10pJ/bit/m^2 and $\mathcal{E}_{amp}$ = 13fJ/bit/m^4.
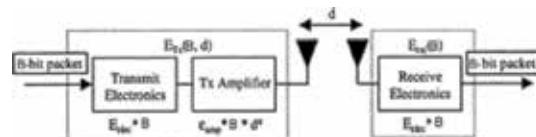
Figure 4.
The radio energy consumption model.

Propagation environment	Value of δ
Free-space	2
Urban area LOS	2.7–3
Urban area no LOS	3–5
Indoor LOS	1.6–1.8
Buildings no LOS	4–6

Table 2.
Typical values of δ.

The energy dissipated for receiving B-bit message E_{Rx} (B) is given by

$$E_{Rx}(B) = E_{elec} * B \tag{5}$$

The E_{elec} by transmitter and receiver circuits are equal and can be set to 50 nJ/bit. The CH and its MNs can communicate provided that the received power will be greater than the sensitivity of the receiver.

The process of data aggregation is same as the concept of feed-forward multi-layer perceptron (MLP) in ANN or fuzzy logic as depicted in **Figure 5**.

The set of input data xi (x1, x2, ... , x_n) are measured by groups of SNs xi in each cluster. The CHs aggregate the data from inputs xi using quantization techniques to process the input data locally, which results in quantized data (q1, q2, ... , qn) before sending them to the fusion center. A weight value wi is assigned to each SN and a bias bk as described in Eq. (5).

$$y_i = f(\mathbf{x}, t) = f\left(\sum_{j=1}^{i} w_{kj}\mathbf{x}_j + b_k\right) \tag{6}$$

The energy consumption due to data aggregation (E_{DA}) is 5pJ/bit.

The CHs can perform sensing (e.g., temperature, pressure, etc.) and compare the readings with data aggregated from other SNs. The sensed signal is analog which will be converted to binary of size D bits using analog-to-digital converter (ADC). If the readings from the SNs fall in the interval [−M, M], their readings will be aggregated, compressed, and sent to upper level CHs or zone gateways which have lower variance (i.e., Da [Da1, Da2] where Da1 ≥ -D and Da2 < D). However, they are identified as outliers since the readings do not match the aggregated confidence interval [Da1, Da2].

The aggregated data are compressed into a single fixed-length packet by applying some sort of transforms on the source bit like discrete wavelet transform (DWT) and discrete cosine transform which achieve signal-to-noise-ratio (SNR) of 20 dB or better to reduce the number of transmitted bits. Relay nodes do not aggregate data as it has less correlation.
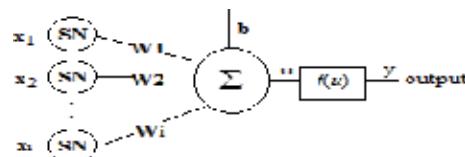
Figure 5.
Data aggregation process in AI model.
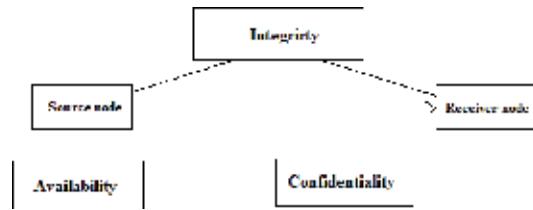
Figure 6.
The attacks on CIA model.

4.3 Security model

Wireless links are susceptible to passive eavesdropping or active impersonation in which the former violates confidentiality by eavesdropping exchanged messages without being detected while later try to inject erroneous messages, drop packets or change routing information.

Symmetric cryptography or group key management is common security scheme for WSNs. Each SNs including the sink needs pairwise cryptographic keys for authentication. The global keys (GKs) are shared among SNs in the network; however, private keys (PKs) are shared between the sink and individual SNs only.

To build the security mechanism, confidentiality, integrity, and availability (CIA) triad model with passive attacks is assumed as shown in **Figure 6.**

5. Proposed protocol

The proposed cross-layer secure MZRP and a hybrid energy-efficient MAC (X-SREM) have three phases: intelligent network configuration phase (INCP), the clustering and rotation of CHs phase (CRCP), and the steady-state phase (SSP). MZRP is a hybrid protocol which divides the sensing field into multi-level hierarchy to solve the minimum transmit power topology control problem faced in multi-hop wireless network. Each level is divided into zones using self-organizing map (SOM) neural network, and zones are further divided into regions which are composed of a group of clusters led by cluster heads (CHs). The BS calculates the proximity of all SNs and those that can communicate directly with the BS will be assigned zone-0 identifier and serve as border gateways (BG); and the CHs in zone-1 forward their data directly to the zone-0 gateways. The CHs in the lower level zones choose the least cost multi-hop paths to forward their data to the BS using multihop.

5.1 Intelligent network configuration phase (INCP)

The initial phase is the network configuration that begins with "hello" message from the BS with the TDMA schedule to all SNs to advertise themselves. Based on the distance, the neural network clustering using SOM is performed in the BS. The input layer of the neural network is the three-element vectors (3×100 matrix) which define the x- and y-coordinates of the SNs and the distance to the BS. The SOM layer is 1-by-10 hidden layers of neurons to classify SNs to zone in a 2D-grid as illustrated in **Figure 7** and described in Algorithm 1.
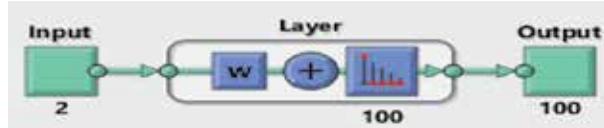
Figure 7.
Self organizing map network diagram.

ALGORITHM 1: INCP to create zones.

Input: N, number of sensors; Z, number of zones; S(x,y), locations of sensors in 2D plane; d, distance to the BS

Output: the zone ID
function Zone (N)
// Perform (SOM) to partition the field into zones using Selforgmap
1: idx (Z)←Selforgmap (x,y,d)
//Assign zone ID to each node
2: **for** each node Nj where j = 1, 2,.., N
3:     S(j).zone ← idx(j)
4: **end**
5: return (Zone)
6: **end** function

5.2 Clustering and rotation of CH phase (CRCP)

Once SNs are assigned to zones, the BS runs ML algorithm (e.g., k-medoids) to determine k CHs in each zone that minimizes the cost function described in [17].

$$\text{cost} = \beta \times f1 + (1 - \beta) \times f2 \tag{7}$$

$$f1 = \max_{i=1, 2, ..k} \sum_{\forall n_i \epsilon C_k} d(n_i, C_k)/|C_k| \tag{8}$$

$$f2 = \sum_{i=1}^{N} E(n_i)/\sum_{j=1}^{k} E(C_k) \tag{9}$$

The BS broadcasts "hello" message again to all SNs which includes information of the elected CHs. Once the CHs are selected, the clusters are fixed for the entire network lifetime and members are sorted in a list to become CHs based on the distance from the BS if the energy level of CHs fall below a threshold value defined as f2/k. The zones created around the BS based on the distance using SOM and CHs using k-medoids are shown in **Figure 8(a)** and data aggregation in **Figure 8(b)**.

5.3 Steady-state phase (SSP)

The SSP breaks the data transmission into frames using hybrid TDMA/CSMA where member nodes (MNs) send the data to their CHs in the time slots allocated to them. Hence, SNs have periodic sleep with duty-cycle equals $T_A/(T_A + T_S)$ where T_A and T_S are the active and sleep time, respectively. This is repeated periodically to save SN's energy.

The CHs aggregate the data and send them to the BS or ZG using carrier sense multiple access (CSMA) and fixed spreading codes. Based on traffic, CHs switch between TDMA and CSMA to transmit the data. For instance, when the energy
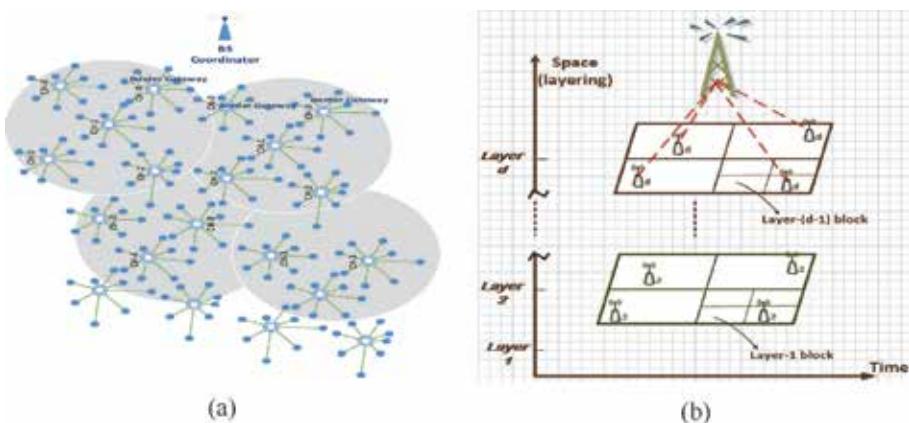
Figure 8.
Zone formation (a) network topology (b) data aggregation.

threshold is reached, the SNs with highest energy will be elected as CHs and this process continues to manage the routing using the AI algorithm in entirely distributed fashion. The round structure is shown in **Figure 9**.

The flow chart of MZRP is shown in **Figure 10**.
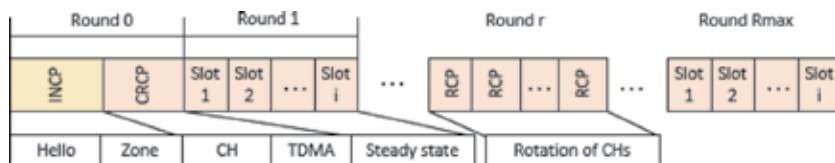
Figure 9.
Round structure in MZRP network model (a) deployment (b) connectivity.
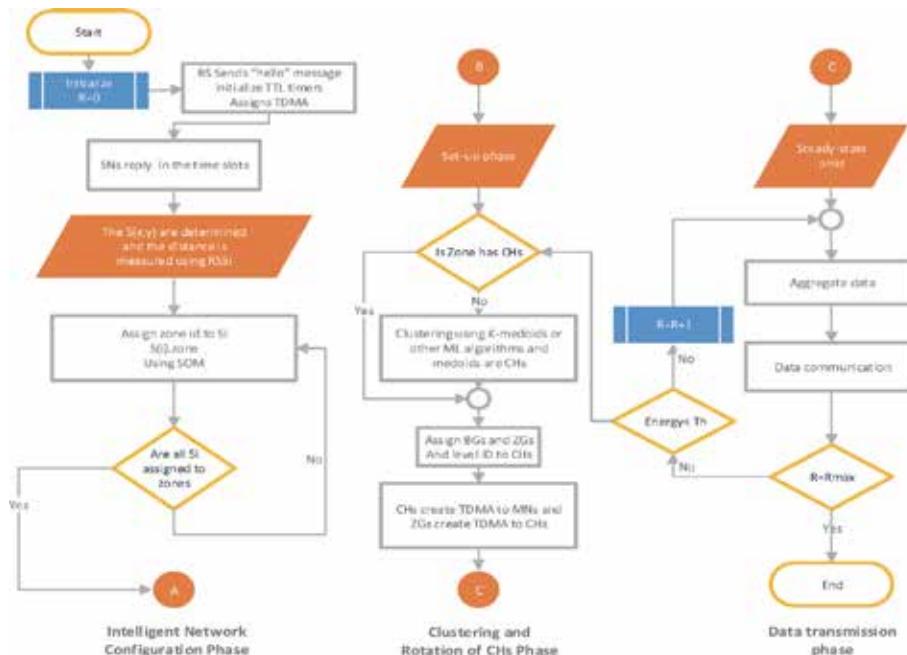
Figure 10.
Three phase of MZRP.
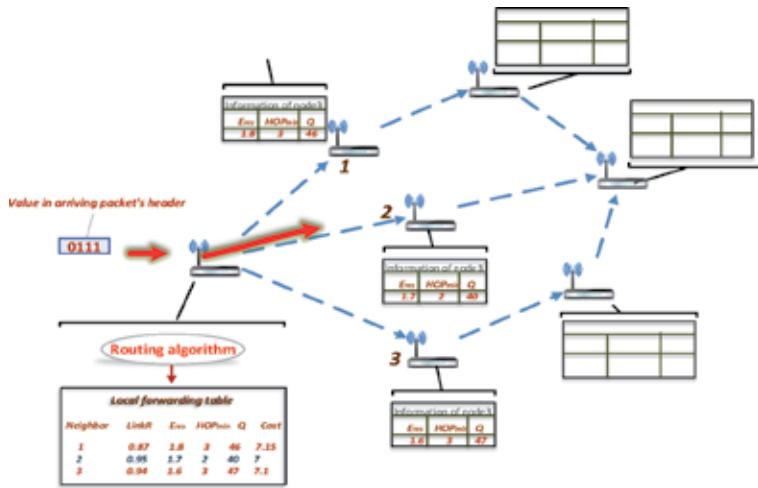
Figure 11.
Route selection in inter-CHs for each zone.

A single-hop inter-cluster communication happens between the MNs and their CHs and between CHs and BS as well if the distance between the BS and the CHs is less than their communication range, otherwise data are relayed from lower-level CHs through intermediate CHs at upper levels until it reaches the sink using multi-hop with minimum cost based on zone ID, remaining energy and link quality with minimum hop-count as shown in **Figure 11**.

The cost of a link (i,j) is given by

$$c_{i,j} = e \times \frac{E_{res,j}}{E_{init,j}} + q \times \frac{Q_{empt,j}}{Q_{init,j}} + l \times R_{i,j} \qquad (10)$$

where E_{res} is the initial energy E_{init} minus the energy consumption $(E_{Tx} + E_{Rx} + E_{DA})$ defined before and e, q, and l are constants. The route selection is described in Algorithm 2.

ALGORITHM 2: Optimal next-hop route

Require: ni, CH neighbors; ZID, zone IDs; HOP, hop count; Eres, remaining energy; R, link quality, Q, queue

Ensure: NH, next-hop ID
1: **while** (the zone ID of $n_i \leq$ zone ID of CH)
2: **repeat**
//calculate the link $c_{i,j}$ for all neighbors of CHs
3: **for** all j ϵ n_i do
4:     $c_{i,j} \leftarrow$ use eq (10)
5: if $E_{res,j} - c_{i,j} < E_{thr,j}$
6:     rotate the CHs
7: end if
8: **endfor**
9: **endwhile**
//list of ni is created
10: j← top of the list
11: **while** (not end of the list ni)

12:    if (HOP next element equals HOP of the CH) then
13:      add j to ni
11:    end if
14: j← next element in the list
15: **endwhile**
16: sort n_i according to $c_{i,j}$ in descending order NH
17: NH← top of the list n_i

Now, logical key tree is produced using topology key hierarchy (TKH) algorithm [25] based on the model shown in **Figure 12**. The algorithm is described below.

The total number of messages in d degree key tree is $\sum_i^{h-1}(di-1)-1$ [26].

The total cost of key update is sum of cost of revoked subtree (RS) (tree$_i$) plus the sum of cost of subtrees $\notin$ RS given by

$$c_{tot} = \sum_{tree_i \in RS} Ctree_i + \sum_{tree_i \notin RS} Ctree_j \tag{11}$$

ALGORITHM 3: Secure routing and aggregation in MZRP

Require: network topology

Ensure: key tree

Function keytree (G, V_m)
  for each tree **do** {
    if (|tree$_i$| = 1) // the BGs
    {attach the root$i$'s PK to the GK}
    elseif (|tree$_i$| ≥ 2)
    {generate ZG key (TK) and attach the root$i$'s PK to the GK}
    current_parent← root$_i$
    for each non-leaf node in tree$_i$ **do** {
      set {sibling set} ← children of current_parent
      if (CH set = 1)
      {attach the sibling's PK to the ZG key (TK$_i$}
      elseif (|sibling set | ≥ 2)
      {generate the CH key SK (Sibling Key) and attach it to the TKi, attach all PKs to the SK}
      current_parent← unvisited non-leaf node in tree$_i$
      endif
    └ endfor
    endif
  └endfor return keytree
└end function

The model in **Figure 12** is a novel mechanism for security in WSNs to defend against aggregation and routing attacks. The subtrees are rooted at T1, T2, and T3 or BGs.

The sink uses a public function F to generate a key chain K_i where $K_i = F(K_i + 1)$. Each SN had stored an initial key K_0 before its deployment where $K_0 = F_n(K)$. Hence, the sink's first message will be encrypted using $K_1 = F_{n-1}(K)$ and transmitted. Once they have been received by all SNs, the sink reveals K_1 and each SN computes $F(K_1) = F(F_{n-1}(K))$ to validate this matching its initial key (i.e., K_0). Similarly, the consecutive keys can be revealed until $K_n = K$ is reached.

Further processing on the sensed data and when decision has to be made, fog nodes have data base storage to store the data generated temporarily and perform AI/ML to make decision on what to do next and send commands to the BS to
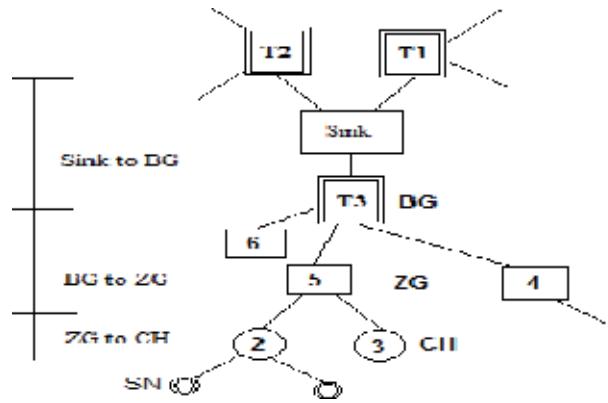
Figure 12.
Secure multicast routing and data aggregation.

instruct it what action needs to be taken. The big data will be sent to the cloud for storage and higher power computation. **Figure 13** illustrates the relationship between IoT, AI/ML and Big data.
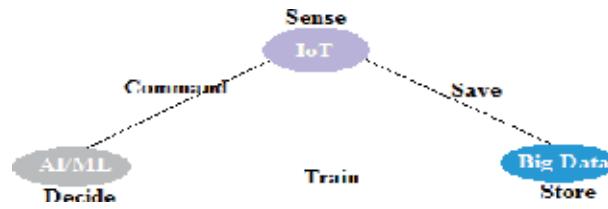
Figure 13.
The IoT, AI/ML and big data relationship.

6. Simulation

MATLAB R2019a is used to evaluate the proposed cross-layer MZRP and a hybrid MAC protocol based on the model described in Section 4.1. The BS performs SOM to create the zones and k-medoids to elect CHs. Other simulation parameters are listed below.

Description	Parameter	Value
Energy in electronics	E_{elec}	50 nJ/bit
Energy of path loss parameters	$\mathcal{E}_{fs,adj}$	10pJ/bit/m^2
	$\mathcal{E}_{mp}$	0.0013pJ/bit/m^4
Initial energy	E_0	0.5 J
Data aggregation	E_{DA}	5nJ/bit/signal
Message size	B	4000 bits
Optimal probability	P_{opt}	0.05

The average remaining energy in MZRP is higher compared to dual-hop and HT2HL as shown in **Figure 14(a)**.
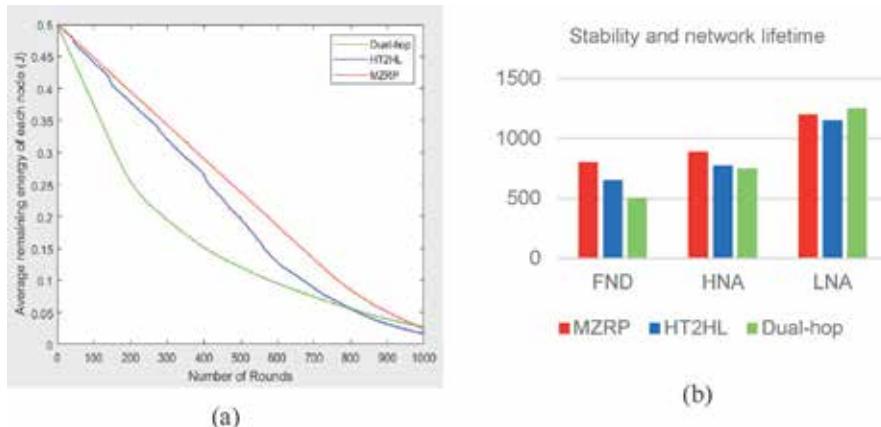
Figure 14.
The average remaining energy (a) and network lifetime (b).

The simulation of MZRP, dual-hop, and HT2HL are repeated, and the average values for the deaths of first node, half nodes, and last node (FND, HND and LND) are shown in the histogram of **Figure 14(b)**. Obviously, MZRP outperforms dual-hop and HT2HL in terms of network lifetime. Hence, it is more energy efficient.

The MAC protocol is responsible for active/sleep schedule. The average power consumption of a SN can be approximated by the following expression:

$$P_{avg} = \left[\frac{T_A \times I_A}{T_{cycle}} + I_S \right].V \qquad (12)$$

where T_{cycle} is the cycle period (i.e., $T_A + T_S$), I_S is the current drawn in the sleep mode, and V is the battery voltage [27].

Figure 15(a) compares the remaining energy in CSMA and TDMA for different numbers of SNs accessing the channel. TDMA is better when the number of SNs is increased as it offers collision free media. Hence, SNs transmit in the time slot and turn their transceivers off. Thus, SNs have less active time in TDMA compared to CSMA as shown in **Figure 15(b)**.

The TKH cryptographic algorithm [25] is used in the MZRP, and the key management is static (i.e., no update) since re-clustering does not happen until the threshold value of the energy of CHs are reached. As the topology will have changed as new CHs are elected, the keys will be updated. Hence, dynamic key management
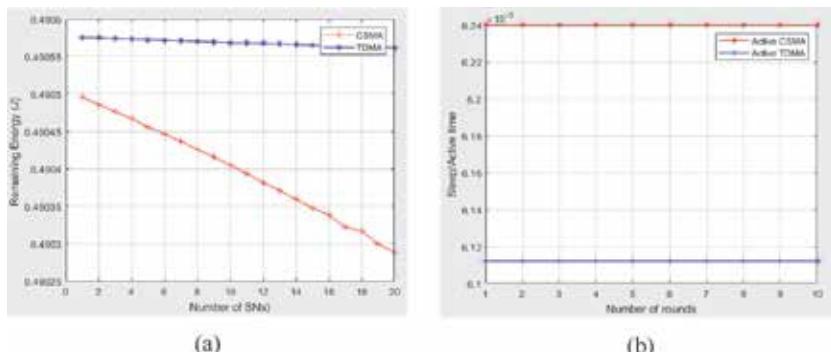
Figure 15.
Performance comparison (CSMA/TDMA).
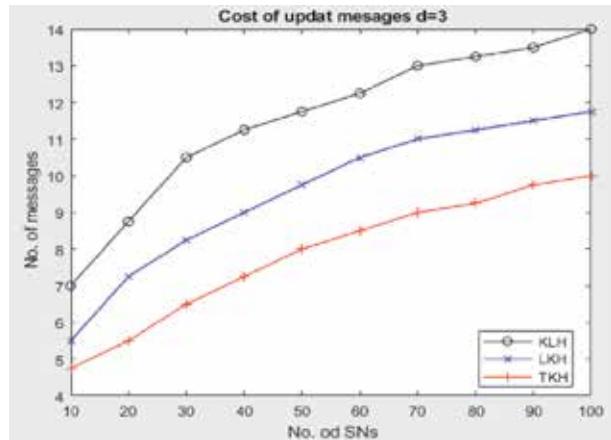
Figure 16.
Comparison of key management algorithms.

is required. The TKH algorithm has low cost compared to logical key hierarchy (LKH) and key link tree (KLT) [26] as shown in **Figure 16**.

Clearly, the TKH has the least key update messages; hence, energy consumption of SNs using TKH is less.

7. Conclusion

The cross-layer energy efficient multi-hop zone-based routing MZRP and hybrid MAC (X-SREM) is proposed in this paper. The AI/ML are at the heart of our design where SOM is used to divide the area into zones and k-medoids algorithm is used for clustering and ANN is used to aggregate data by CHs. MZRP protocol extends the network lifetime; hence, its performance of is better than dual-hop and HT2HL. The load balancing is achieved through rotation of CHs within clusters when the threshold value of the energy level is 50% of the initial energy since this has been tested to be the optimum threshold for reclustering. The fog nodes provide data analytics near the network edge and relay big data to cloud for further processing and storage. Enhanced lifetime is our future work to improve MZRP.

Author details

Saad Alharthi[1*], Princy Johnson[1] and Martin Randles[2]

1 Department of Electronics and Electrical Engineering, Liverpool John Moores University, Liverpool, UK

2 Department of Mathematics and Computer Science, Liverpool John Moores University, Liverpool, UK

*Address all correspondence to: ieeeeng@yahoo.com

IntechOpen

References

[1] Alharthi S et al. IoT/CPS ecosystem for efficient electricity consumption. In: Tenth International Green and Sustainable Computing Conference (IGSC), Alexandria, VA, USA. 2019. pp. 1-7

[2] Gupta A, JHA RK. A survey of 5G networks: Architecture and emerging technologies. IEEE Access. 2015;**3**: 1206-1232

[3] Mouradian C et al. A comprehensive survey fog computing: State-of-the-art and research challenges. IEEE Communication Surveys and Tutorials. 2018;**20**(1):416-464

[4] Whitmore A, Agarwal A, Da Xu L. The internet of things: A survey of topics and trends. Information Systems Frontiers. 2015;**17**:261-274

[5] Zhou F et al. Energy-efficient NOMA heterogeneous cloud radio access networks: Enabling techniques and challenges. IEEE Network. 2018;**32**(2): 152-160

[6] Taneja M, Davy A. Resource aware placement of data analytics platform in fog computing. Procedia Computer Science. 2016;**97**:153-156

[7] Heinzelman W, Chandrakasan A, Balakrishnan H. Energyefficient Communication Protocol for Wireless Microsensor Networks. Maui, HI, USA: Proceedings of the IEEE Hawaii International Conference on System Sciences; 2000. pp. 1-10

[8] Heinzelman W, Chandrakasan A, Balakrishnan H. An application specific protocol architecture for wireless microsensor networks. IEEE Transactions on Wireless Communications. 2002;**1**(4):660-670

[9] Manjeshwar A, Agrawal DP. TEEN: A protocol for enhanced efficiency in wireless sensor networks. In: Proceedings of the 1st International Workshop on Parallel and Distributed Computing Issues in Wireless Networks and Mobile Computing. San Francisco, USA; 2001

[10] Manjeshwar A, Agrawal DP. APTEEN: A hybrid protocol for efficient routing and comprehensive information retrieval in wireles sensor networks. In: Proceedings of the 2nd International Workshop on Parallel and Distributed Computing Issues in Wireless Networks and Mobile Computing. San Francisco, USA; 2002

[11] Younis O, Fahmy S. HEED: A hybrid, energy- efficient, distributed clustering approach for ad hoc sensor networks. IEEE Transactions on Mobile Computing. 2004;**3**(4):366-379

[12] Smaragdakis G, Matta I, Bestavros A. SEP: A stable election protocol for clustered heterogeneous wireless sensor networks. In: 2nd International Workshop on Sensor and Actor Network Protocols and Applications (SANPA 2004). 2004. pp. 1-11

[13] Qing L, Zhu Q, Wang M. Design of a distributed energy-efficient clustering algorithm for heterogeneous wireless sensor networks. Computer Communications. 2006;**29**:2230-2237

[14] Alharthi SA, Johnson PA. Threshold sensitive heterogeneous LEACH protocol for wireless sensor networks. In: 24th Telecommunications Forum (TELFOR). Belgrade; 2016. pp. 1-4

[15] Abushiba W, Johnson P, Alharthi S, Wright C. An energy efficient and adaptive clustering for wireless sensor network (CH-leach) using leach protocol. In: 13th International Computer Engineering Conference (ICENCO). Cairo; 2017. pp. 50-54

[16] Rekha PA. Energy efficient of inter cluster multihop routing protocol for wireless sensor network. In: International Conference on Interactive and Advanced Communication Technology (ICIACT2016). 2016. pp. 1-8

[17] Latiff NA, Tsimenidis C, Sharif B. Performance comparison of optimization algorithms for clustering in wireless sensor networks. In: IEEE International Conference on Mobile Adhoc and Sensor Systems. Pisa; 2007. pp. 1-4

[18] Palan NG, Barbadekar BV, Patil S. Low energy adaptive clustering hierarchy (LEACH) protocol: A retrospective analysis. In: International Conference on Inventive Systems and Control (ICISC), Coimbatore. 2017. pp. 1-12

[19] Loscri V, Morabito G, Marano S. A two-levels hierarchy for low-energy adaptive clustering hierarchy (TL-LEACH). In: VTC-2005-Fall. 2005 IEEE 62nd Vehicular Technology Conference. Dallas, TX, USA; 2005. pp. 1809-1813

[20] Kumar S et al. Resource efficient clustering and next hop knowledge based routing in multiple heterogeneous wireless sensor networks. International Journal of Grid and High Performance Computing (IJGHPC). 2017;**9**(2):1-20

[21] Khurnan S, Rekha KA. Energy efficient of inter cluster multihop routing protocol for wireless sensor network. In: International Conference on Interactive and Advanced Communication Technology (ICIACT2016). 2016. pp. 1-8

[22] Al-Sodairi S, Ouni R. Reliable and energy-efficient multi-hop LEACH-based clustering protocol for wireless sensor networks. Sustainable Computing: Informatics and Systems. 2018;**20**:1-13

[23] Alharthi SA, Johnson P, Alharthi MA. IoT architecture and routing for MV and LV smart grid. In: Saudi Arabia Smart Grid (SASG). Jeddah; 2017. pp. 1-6

[24] Guo W, Zhang W. A survey on intelligent routing protocols in wireless sensor networks. Journal of Network and Computer Applications. 2014;**38**: 185-201

[25] Son J, Lee J, Seo S. Energy efficient group key management scheme for wireless sensor networks. In: 2nd International Conference on Communication Systems Software and Middleware. Bangalore; 2007. pp. 1-9

[26] Li G, Wang Y, He J. Efficient group key management scheme in wireless sensor networks. In: Third International Symposium on Intelligent Information Technology and Security Informatics. Jinggangshan; 2010. pp. 411-415

[27] López M et al. Software management of power consumption in WSN based on duty cycle algorithms. In: Eurocon. Zagreb; 2013. pp. 399-406

A Mobile Ad Hoc Network Routing Protocols: A Comparative Study

Alagan Ramasamy Rajeswari

Abstract

Mobile Ad hoc NETworks (MANET), are complex and distributed networks that are dynamic. Which are infrastructure less and multi-hop in nature. The communication of a node can be either direct or through intermediate nodes without a fixed and dedicated infrastructure. Hence it is necessary to design an efficient routing protocol for ad hoc network which can address the issues of MANET efficiently. In ad hoc, routing algorithms are classified into nine categories namely: source-initiated (reactive), table-driven (proactive), hybrid, hierarchical, multipath, multicast, location-aware, geographical-multicast and power-aware. This paper presents a survey and to review a comparative study about various routing protocols under each of these categories. Additionally, brief discussions about major routing issues are addressed. This survey paper focuses on the taxonomy related to ad hoc routing techniques and compares the features of routing protocols.

Keywords: ad hoc networks, routing protocols, survey, wireless network

1. Introduction

A wireless network can work under two modes namely infrastructure and infrastructureless. In the "ad hoc" topology, the user does not rely on fixed infrastructure where the nodes are self-configured and self-managed. On the other hand in "infrastructure "topology, the nodes are under the control of a centralized authority called base station. Wireless multi-hop networks, also known as ad hoc networks have been used in many applications like military, disaster relief communications and emergency. An ad hoc network is a self-organizing multi-hop wireless network, which is independent neither on fixed infrastructure nor on predetermined connectivity. It is a collection of nodes, which communicate with each other using radio transmissions. In ad hoc network, there is no base station to act as router. The intermediate nodes will act as a router; source node will use these nodes for routing their message. Thus, each and every of the node forwards packets on behalf of other nodes until the packet is received by the destination from its sender. Therefore, data should be forwarded from source to destination through multiple hops. Ad hoc networks rely on multi-hop transmissions among the nodes in the same channel. Nodes communicate with each other through the intermediate nodes. So, the efficient performance and availability of each node is important in ad

hoc network environment. Hence an efficient routing protocol is required to enhance the communication in MANET.

Thus routing becomes a major challenging task in MANET. In this paper, a review about the technologies, characteristics, advantages and disadvantages of the routing protocols in ad hoc network are provided.

The paper continues as follows, Section 2 summarizes the issues involved in ad hoc routing protocols. In Section 3 the routing protocols are organized as follows.

- Reactive (on-demand) (Section 3.1).

- Proactive (table-driven) (Section 3.2).

- Hybrid (Section 3.3).

- Hierarchical (Section 3.4).

- Multipath (Section 3.5).

- Multicast (Section 3.6).

- Geographical (location-aware) (Section 3.7)

- Geographical multicast (Section 3.8).

- Power-aware (Section 3.9).

Finally, Section 4 provides the conclusion for this paper.

2. Issues with ad hoc routing protocols

Due to the highly dynamic nature of mobile ad hoc network, it results in frequent and unpredictable changes in network topology and hence makes routing among the mobile nodes as a complex and difficult task. The challenges and complexities together with the importance of routing protocols make the routing process, as the most active and innovative research area in the MANET domain. The issues in routing techniques includes the large area of flooding, greedy forwarding, flat addressing and widely distributed information, large power consumption, interference and load balancing [1] (**Table 1**).

3. Routing protocols

In this section, the categories of routing protocols are elaborated in detail manner and the overall performance of the routing protocols in ad hoc network are evaluated for each protocol under various routing categories by considering the following parameters such as route metrics, time complexity, computation complexity and route structure. One of the main features of routing protocols in ad hoc network is the routing metric, which is used to select the best route for forwarding packets. Time complexity (TC) is defined as the time required for the number of steps to perform a protocol operation, communication complexity (CC) is defined as the number of messages exchanged in performing a protocol operation and

Issues	Protocol	Approach
Large area of flooding—flooding is a routing technique used to forward packets from the source to destination during the route discovery phase or in a recovery phase.	1. Distance routing effect algorithm for mobility (DREAM) [2]	Flooding area is reduced by limiting number of neighbors that can forward a route request message.
	2. Location aided routing (LAR) [3]	Nodes location information is used for routing the packets. Limiting the flooding area into "request zone".
	3. Location based multicast (LBM) [4]	Similar to LAR, limiting the flooding area into "forwarding zone"
	4. Geographical distance routing (GEDIR) [5]	Greedy forwarding approach is used.
	5. Temporally-ordered routing algorithm (TORA) [6]	DAG is constructed rooted at destination and ordered routing scheme is used
	6. Geographical GRID (GeoGRID) [7]	The process of portioning the geographical area of the network into smaller areas called grids.
	7. Geographical TORA (GeoTORA) [8]	Uses any-cast any group-cast forwarding approach
	8. Zone routing protocol (ZRP) [9]	Overlapped zone are created based on the separation distance between the mobile nodes. Peripheral nodes are selected to forward the control packets within the zone.
	9. Mobile just-in-time multicasting (MOBICAST) [10]	Routing area is divided into two parts namely: a delivery zone and forwarding zone.
Greedy forwarding—greedy forwarding (GF) is one the routing technique that relies on only single path from the source to its destination which is discovered. By using GF, the major challenges encounter is defined as 'GF empty neighbor set problem'. The forwarding process reaches a dead end, when a node cannot find any neighbor which is closer to the destination then itself.	1. FACE routing protocol [11]	Unit graph approach is utilized: two nodes communicated with each other if the Euclidean distance between them is less than some fixed amount.
	2. Geographical routing without location information (GRLI) [12]	Extended ring search: this search process continues until a node closes to destination is identified else if not found, the search is extended until a predetermined a TTL
	3. Bounded Voronoi greedy forwarding (BVGF) [13]	Greedy routing decision is based on the location of the direct neighbors of each node.
	4. Greedy distributed spanning tree routing (GDSTR) [14]	Hull tree approach is used
	5. Greedy perimeter stateless routing (GPSR) [15]	Greedy forwarding and perimeter forwarding approach are used
Flat addressing and widely-distributed information—In a MANET, due to the distribution of mobile nodes over the network and the restriction in transmission range of each node's may cause some nodes to have poor knowledge about the network.	1. Grid location service (GLS) [16]	Based on distributed location servers (DLS) avoids the congestion in the node.
	2. Dynamic address routing (DART) [17]	Utilizes dynamic address scheme that ensures the scalability
	3. GPS ant-like routing algorithm (GPSAL) [18]	Based on ant colony optimization. A specifically defined node namely ant node is responsible for collecting and

Issues	Protocol	Approach
		forwarding the location information around the network.
	4. Augmented tree-based routing (ATR) [19]	Based on the structured address space.
Large power consumption—in a MANET, routing techniques depends upon the battery power of the node. Thus more power consumption will increase the overhead in the transmission.	1. Infra-structure AODV for infrastructure ad-hoc networks (ISAIAH) [20]	1. Modified forwarding approach: selecting the routes that pass through power base stations (PBSs) instead of through mobile nodes. Thus the amount of power utilized by each node can be reduced. 2. Nodes are allowed to enter the power-saving mode for a fixed time period that will decrease the power consumed by the node.
	2. Power-aware routing optimization protocol (PARO) [21]	New forwarding node–redirectors are added on the routing path to reduce the transmission power of the intermediate nodes along the original path. The objective of PARO to increase the path length to reduce the total transmission power.
	3. Dynamic source routing power-aware (DSRPA) [22]	The routing metric battery freshness is considered in routing to achieve connectivity for the longest period of time.
	4. Power-aware multi-access protocol with signaling ad hoc networks (PAMAS) [23]	The battery usage is controlled based on the frequency of a node's activities.
Inference and load balancing—Interference is a major problem factor that affects the performance of wireless networks. Routing in a wireless network is challenging due to the unpredictable nature of the wireless medium and due to the effect of interference on wireless link properties.	1. Source Routing for Roofnet (SrcRR) [24]	Expected transmission counts (ETX): is an interference-aware link-based routing metric that continuously measures the link loss rate in both directions between each node about its neighbors using periodic broadcasts. Link cost is estimated considering the number of retransmission attempts.
	2. Link quality source routing (LQSR) [25]	Weighted cumulative expected transmission time (WCETT). WCETT is the sum all links costs (ETT) along the path and bottleneck channel which has the maximum sum of ETT.
	3. Load-balancing routing for mesh networks (LBRMN) [26, 27]	Metric of interference and channel-switching (MIC) and isotonic property.
	4. Interference-aware load-balancing routing (IALBR) [28]	Routing metric load value: defined as the load at node itself and its next hop node load.
	5. Load-balancing curveball routing (LBCR) [29]	Modified route metric based on greedy routing scheme.

Table 1.
Comparison of various issues in routing protocols and solutions to handle the issues.
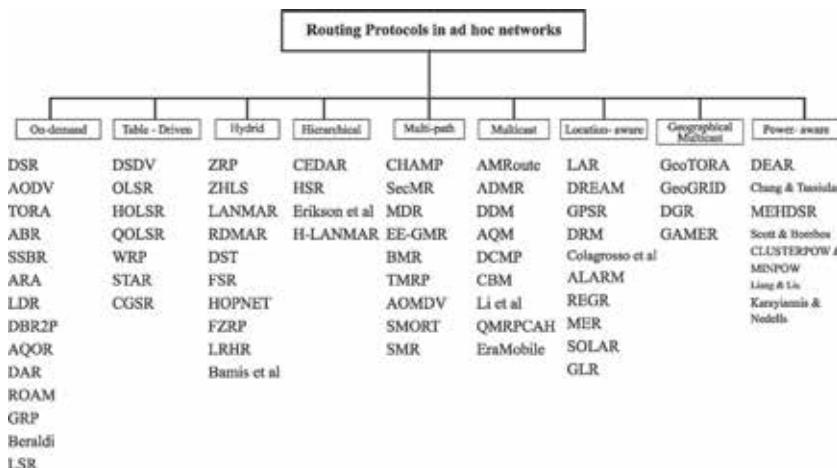
Figure 1.
Categories of ad hoc routing protocols.

finally, route structure defines whether the structure and address scheme are flat or hierarchical. **Figure 1** depicts the categories of ad hoc routing protocols.

3.1 Reactive (on-demand) routing protocol

In reactive routing protocols, a node initiates a route discovery, only when it wants to send packet to its destination. They do not maintain or constantly update their route tables with the latest route topology.

Therefore, the communication overhead is reduced but the delay is increased due the on-demand route establishment process.

Dynamic source routing (DSR): DSR is a primary on-demand routing protocol proposed by Johnson et al. [30] DSR is a most widely known protocol that relays on source routing mechanism. The network bandwidth overhead is reduced by transmitting the routing message on-demand and battery power is harvested on the nodes since each of the nodes has to transmit the control packets whenever needed.

Adhocon-demand distance vector (AODV): Perkins et al. [31] proposed AODV to provide loop- free routes even under the condition of repairing the failure routes. The Time to Live (TTL), prevents the unnecessary forwarding of packets by a node hence reduces control overhead. Since, the performance depends on the bandwidth and end-end delay, so the route cache mechanism is not implemented in this protocol.

Temporally ordered routing algorithm (TORA): Park and Corson developed TORA [6] an adaptive and scalable routing algorithm. TORA is based on "link reversal" algorithm. This protocol is proposed to operate in a highly dynamic mobile wireless network environment. A directed acyclic graph (DAG) rooted at a destination is constructed by using a height as a metric.

Associativity based routing (ABR): Tohn [32] developed the ABR a simple and width efficient distributed routing algorithm. ABR exploits route stability as the criteria in selecting a best route. ABR algorithm uses a mechanism called associativity ticks to determine and maintain a "degree of associativity". The protocol is loop-free, no deadlock condition, no duplicate of packets.

Signal stability-based adaptive routing (SSBR): SSBR, by Dube et al. [33] is a distributed adaptive routing protocol designed for ad hoc network by considering the signal strength and location stability as the routing criteria. Thus, the final path

from source to destination consists of only strong link. If multiple paths are available, then the destination selects one route among them.

The ant colony based routing algorithm (ARA): Gunes et al. [34] proposed an innovative mechanism for on-demand, multi hop ad hoc routing, based on swarm intelligence and the ant colony meta heuristic. ARA is designed with a primary objective to reduce the overhead without any direct link among the participants the complex optimization and collaboration problem are solved by this type of algorithm.

Labeled distance routing (LDR): Luna-Aceves et al. [35] presented an on-demand, loop free routing protocol. LDR utilizes distance labels to ensure loop free path in the network rather than using sequence number as other routing algorithms. LDR exploits a RouteRequest, RouteReply and RouteError packet as.

Dynamic backup routes routing protocol (DBR^{2}P): DBR^{2}P, an on-demand routing protocol by Wang and Chao [36]. The special unique feature about DBR^{2}P is, it does not require any routing table as other routing protocols.

AdhocQoS on-demand routing (AQOR): AQOR, an on-demand routing protocol enabling QoS support in terms of bandwidth and end-end delay is developed by Xue and Ganz [37]. AQOR mechanism estimates the bandwidth and end-end delay requirements and exploits these metrics to determine accurate admission control and resource reservation decision. TTL, prevents the unnecessary forwarding of packets by a node hence reduces control overhead.

Distributed ant routing (DAR): DAR, a distributed algorithm developed by Rosati et al. [38] DAR is based on the ant behaviour in colonies. The goal of DAR is to reduce the computation complexity. Each node maintains a routing table. Forward ants are used to find new route. A node selects the next hop node based on weighted probabilities.

Routing on-demand a cyclic multipath (ROAM): Raju and Garcia-Luna-Aceves [39] proposed ROAM, based upon the directed acyclic graphs (DAG).

Gathering based routing protocol (GRP): GRP by Ahn [40] collects network information during route discovery process. The source node uses the network information collected during route discovery process to forward the packets even if the current route is failed. The source node computes the optimal path based on the collected network information. Then, through the optimal path data packets are forwarded.

Hint based probabilistic protocol: Beraldi et al. [41] in this protocol, the nodes of the network uses a set of meta-information defined as hints to discover a route to the destination. This protocol has lower control overheads.

Preemptive routing in ad hoc networks: Goffe et al. [42] developed a routing algorithm. The algorithm initiate the route discovery process to discovery an alternative route before the probable current route failure.

Labeled successor routing (LSR): Rangarajan and Garcia-Luna-Aceves [43] presented LSR. According, to authors view many on-demand protocols are built on top of AODV, by exploiting sequence number. **Table 2** illustrates the comparative analysis of reactive routing protocols.

3.2 Proactive (table driven) routing protocol

In proactive routing, each node has one or more tables that contain the latest information of the routes to any node in the network.

Destination sequenced distance vector routing (DSDV): DSDV, based on BellmanFord routing mechanism is a table-driven routing protocol was developed by Perkins and Bhagwat [44].

Optimized link state routing (OLSR): Clausen et al. [45] proposed the OLSR, a proactive routing protocol based on the link state routing.

OLSR with quality of service (QOLSR): QOLSR, proposed by Munaretto and Fonseca [46] is based on the traditional OLSR.

Protocol	RS	Beacons	Route metrics	Route repository	Route reconfiguration strategy	CC	TC
DSR	F	No	Hop-count	RC	SN and new route	O(2n)	O(2d)
AODV	F	Yes	Hop-count	RT	SN and new route	O(2n)	O(2d)
TORA	F	No	Hop-count	RT	Link reversal and route repair	O(2n)—during route discovery O(2a)—during route maintenance	O(2d)
ABR	F	Yes	Degree of association stability	RT	Local broad cast query	O(n + y)—during route discovery O(x + y)—during route maintenance	O(d + z)—during route discovery O(l + z)—during route maintenance
SSBR	F	Yes	Strong signal strength	RT	SN and new route	O(n + y)—during route discovery O(x + y)—during route maintenance	O(d + z)—during route discovery O(l + z)—during route maintenance
GoFF et al.	F	Yes	Signal strength	RT	New path discovery before route failure	O(2n)	O(2d)
AQOR	F	Yes	Bandwidth	RT	Initiate from destination	O(2n)	O(2d)
ARA	F	No	Hop-count	RT	Alternate route or back track until new route is identified	O(n + r)—during route discovery O(n + a)—during route maintenance	O(d + p)
ROAM	F	No	Hop-count	RT	Erase route and start a new search to get new route	O(|e|)—during route discovery O(6G$_a$)—during route maintenance	O(d)—during route discovery O(x)—during route maintenance
DAR	F	No	Weighted probabilities	Stochastic RT	New route by forward ant	O(2n)	O(2d)
LSR	F	No	Relay sequence label	RT	SN and new route/local repair	O(2n)	O(2d)
GRP	F	No	Hop-count	RC	Route backup	O(2n)	O(2d + 1)
LDR	F	No	Hop-count	RT	SN and new route/local repair	O(2n)	O(2d)
DBR2P	F	No	Hop-count	None	Local repair	O(2n)	O(2d)
Beraldi et al.	F	Yes	Hint value	RC	Local broadcast query	O(2n)	O(2d)

RS = routing structure; H = hierarchical; F = flat routing repository; RC = route cache; RT = route table; RM = route metric; SP = shortest path; CC = communication complexity; TC = time complexity; n = number of nodes in the network, d = diameter of the network, |e| = number of edges on the network, g = maximum degree of the router, z = diameter of the directed path where the REPLY packet transits, l = diameter of the affected network segment, y = total number of nodes forming the directed path where the Reply packet transmits, p = diameter of direct path of the reply, x = number of clusters.

Table 2.
Comparative analysis of reactive routing protocols.

Hierarchical proactive routing mechanism for mobile ad hoc networks (HOLSR): Villasensor-Gonzalez et al. [47] proposed HOLSR protocol, which was developed based on OLSR by organizing node in a hierarchical structure to overcome the inefficiency faced by the flat routing protocol in exploiting the nodes with higher source like bandwidth, transmission range etc.

Wireless routing protocol (WRP): WRP protocol by Murthy and Garcia-Luna Aceres [48] uses the properties of the distributed Bellman-Ford algorithm. Route is chosen by selecting a neighbor node that would minimize the path cost.

Source tree adaptive routing protocol (STAR): STAR proposed by Garcia-Luna Aceves and Spohn [49]. Using a source tree structure each node defines and store the preferred route to all possible destinations. ORA and LORA are two distinct approaches proposed under STAR protocol. ORA approach is preferred to obtain the optimal path with respect to metric (i.e.) number of hops. With ORA it is possible to obtain feasible paths with fewer packets overhead, but with LORA route do not guarantees to be optimal.

Cluster head gateway switch routing protocols (CGSR): CGSR employs a hierarchical network topology, proposed by Chiag et al. [50] CGSR is based on a distributed algorithm namely least cluster change (LCC). Cluster head is elected by using LCC. LCC algorithm is considered to be stable algorithm for cluster head election. Clustering enables an effective way for channel allocation.

Table 3 describes the comparative analysis of proactive routing protocols.

3.3 Hybrid routing protocols

Hybrid routing protocols are designed with the route discovery mechanism and the table maintenance mechanism features of reactive and proactive respectively.

Protocol	RS	Routing tables	No. of tables	HM	Update frequency	Critical node	RM	CC	TC
DSDV	F	Yes	2	Yes	Periodic	No	Hop-count	O(n)	O(d)
R-DSDV	F	Yes	2	Yes	Probalastic	NO	Hop-count	O(n)	O(d)
OLSR	F	Yes	3	Yes	Periodic	No	Hop-count	O(n)	O(d)
HOLSR	H	Yes	3	Yes	Periodic	Yes, cluster head	Hop-count	O(n)	O(d)
CGSR	H	Yes	2	No	Periodic	Yes, cluster head	Hop-count	O(n)	O(d)
QOLSR	H	Yes	3	Yes	Periodic	No	Delay, bandwidth, hop-count	O(n)	O(d)
WRP	F	Yes	4	Yes	Periodic	No	Hop-count	O(n)	O(d)
GSR	F	Yes	3	No	Periodic with neighbor	No	Hop-count	O(n)	O(d)
STAR	H	Yes	1	No	Only at specific events	No	Hop-count	O(n)	O(d)

RS = routing structure; H = hierarchical; F = flat; CC = communication complexity; TC = time complexity; n = number of nodes in the network; d = diameter of the network.

Table 3.
Comparative analysis of proactive routing protocols.

Hybrid protocol is suitable for ad hoc network where large numbers of nodes are present. The protocols discussed in this section overcome the drawbacks of both proactive and reactive routing protocols such as latency and overhead problems in the network.

Zone routing protocol (ZRP): ZRP proposed by Samer et al. [51] is a hybrid routing protocol. This protocol has features of both proactive and reactive mechanism. In ZRP two different routing approaches are exploited: intrazone routing protocol (IARP) and interzone routing protocol (IERP).

Zone based hierarchical link state routing protocol (ZHLS): ZHLS by Joa-Ng and Lu [52] developed network which is divided into non-overlapping zones based on geographical information.

Landmark ad hoc routing (LANMAR): LANMAR by Pei et al. [53] is a novel routing protocol. LANMAR have combined features of both FSR and Landmark routing. A subnet, set of nodes are grouped together as a single unit are likely to move as a group.

Relative distance micro-discovery ad hoc routing (RDMAR): RDMAR proposed by Aggelous and Tafazoli [54] is loop-free highly adaptive, efficient and scalable protocol. RDMAR consists of two main algorithms: the route discovery algorithm and route maintenance algorithm.

Distributed spanning tree (DST) routing: DST by Radhakrishnan et al. [55] proposed a routing algorithm based on the distributed spanning trees. DST proposes two different routing strategies to determine a route between a source and a destination pair namely: (1) Hybrid tree flooding (HFT) and (2) Distributed spanning tree (DST) shuttling.

Distributed dynamic routing (DDR) algorithm: DDR by Nikaein et al. [56] is a tree based routing protocol. In DDR the trees do not require a root node. In this algorithm the tree are constructed by exchanging the periodic beacon messages among neighbors' nodes.

Fisheye state routing (FSR): FSR based on link state routing algorithm, designed by Pei et al. [57]. FSR maintains the accurate distance and path quality information about the immediate neighboring nodes. FSR are more scalable to large networks.

Hybrid ant colony optimization (HOPNET): Wang et al. [58] proposed a hybrid ant colony optimization (HOPNET) based on nature–inspired algorithm such as ant colony based optimization (ACO) and zone routing.

Fisheye zone routing protocol (FZRP): FZRP presented by Yang and Treng [59] inherits the idea of fisheye state routing in ZRP.

Link reliability based hybrid routing (LRHR): Xiaochuan et al. [60] proposed a novel hybrid routing protocol namely, LRHR. LRHR achieves the dynamic switching between table driven and on demand routing strategies due to the frequent topology changes in the network.

Mobility aware protocol synthesis for efficient routing: Bamis et al. [61] proposed a new stability metric to determine the mobility level of nodes in a network. Based upon this metric the nodes can be classified into different mobility classes in turn determines the most suitable routing techniques for a particular source to destination pair. **Table 4** illustrates the comparative analysis of hybrid routing protocols.

3.4 Hierarchical routing protocols

Hierarchical routing protocols apply clustering techniques to build a hierarchy of nodes. Nodes are organized into groups called zones (or) clusters. Each cluster consists of one or more clusters and gateways. Hierarchical routing protocols are

Protocol	RS	Multiple routes	Beacons	RM	Route repository	Route rebuilding	Critical node	CC	TC
ZRP	F	No	Yes	Through put, end-end delay, packet loss percentage	Intrazone and interzone RT	Route repair at failure point	Yes	Intra-$O(Z_N)$ Inter-$O(N + V)$	Intra-$O(I)$ Inter-$O(2D)$
FSR	F	No	No	Scope range	RT	SN	No	$O(N)$	$O(D)$
LANMAR	H	No	Yes	Hop count	RT at landmark node	SN	Yes	$O(N)$	$O(D)$
RDMAR	F	No	No	Hop count	RT	SN and new route	No	$O(N)$	$O(D)$
SLURP	H	Yes	No	Power consumed	location Cache	SN	Yes	During route discovery Intra-$O(2Z_D)$ Inter-$O(2D)$b During route maintenance Intra-$O(2N/M)$ Inter-$O(2Y)$	Intra-$O(2 N/M)$ Inter-$O(2Y)$
ZHLS	H	Yes	No	End-end delay, packet loss percentage	Intrazone and interzone RT	Location request sent	Yes	During route discovery Intra-$O(N/M)$ Inter-$O(N + V)$ During route maintenance Intra-$O(N/M)$a Inter-$O(N + V)$	Intra-$O(I)$ Inter-$O(D)$
DST	H	Yes	No	Power consumed, hop count	RT	Holding time or shuttling	Yes	Intra-$O(Z_N)$ Inter-$O(N)$	Intra-$O(Z_D)$ Inter-$O(D)$
DDR	H	Yes	Yes	Stable routing	Intrazone and interzone RT	SN	Yes	Intra-$O(Z_N)$ Inter-$O(N + V)$	Intra-$O(I)$ Inter-$O(2D)$
HOPNET	H	No	No	Hop-count	Intrazone and interzone RT	Route repair at failure point	Yes	$O(n)$	$O(D)$
LRHR	F	Yes	Yes	Edge weight	RC,RT	New route discovery	No	$O(n)$	$O(D)$
FZRP	H	No	Yes	Hop-count	Intrazone and interzone RT	Route repair at failure point	Yes	$O(n)$	$O(D)$

Table 4.
Comparative analysis of hybrid routing protocols.

developed with an ability to address scalability issues in ad hoc network environment and to minimize excessive overhead. This on the other side increases the tediousness of the routing techniques used by these protocols.

Core-extraction distributed ad hoc routing (CEDAR): Sivakumar et al. [62] introduced CEDAR, an QoS routing algorithm. In CEDAR a subset of nodes are grouped as the core of the network.

Hierarchical state routing (HSR): HSR is a dynamic, distributed multilevel cluster based hierarchical protocol, proposed by Iwata et al. [63] In HSR clustering schema play a vital role. The primary objective of clustering is to have the efficient utilization of radio channel resource and the reduction of routing overhead, Thus the network performance can be enhanced.

Dynamic address approach: Eriksson et al. [64] introduced a dynamic addressing scheme that can enhance scalability in ad hoc network. Under this scheme a geographical location based dynamic address is added to the nodes permanent identifier.

Hierarchical landmark routing (H-LANMAR): H-LANMAR [65] uses, backbone network mechanism, improve the scalability of the network. In H-LANMAR, nodes in the network are grouped into dynamic multihop clusters. Cluster head is referred as backbone node (BN). In case of backbone failure LANMAR schema is used for packet transmission.

Table 5 illustrates the comparative analysis of hierarchical routing protocols.

3.5 Multipath routing protocols

The multipath routing protocols are designed with primary objectives to provide reliable communication and to ensure load balancing as well as to improve quality of service (QoS) of ad hoc environment. Multipath routing protocols address issues such as multiple paths discovery and maintaining these paths.

Caching and multipath routing protocol (CHAMP): CHAMP protocol, proposed by Valera et al. [66] exploits data caching and shortest multipath routing. The main design goal is to minimize the packet drops that occur due to the frequent route breakages.

Secure multipath routing (secMR): SecMR, secure an on-demand multipath routing protocol is designed by Mavropodi et al. [67]. Many security enhancement techniques are imposed in this protocol to present security attacks of collaborating malicious nodes. A centralized Certifying Authority (CA) issues a certificate to the secret keys.

Energy and mobility aware geographical multipath routing protocols (EM-GMR): Liang and Ren [68] developed energy and mobility aware geographical multipath routing protocol, a fuzzy logic mechanism based multipath routing protocol. According to EM-GMR, while choosing the next hop, a mobile node should

Protocol	Routing tables	No. of routing tables	Update frequency	Hello message	Critical node
HSR	Yes	2	Periodic, within each subset	Yes	Yes, cluster head
CEDAR	Yes	1	On demand	No	Yes
Eriksson et al.	Yes	2	Periodic	No	No
H-LANMAR	Yes	2	Periodic	No	Yes

Table 5.
Comparative analysis of hierarchical routing protocols.

consider the following constraints namely: the remaining battery capacity, mobility and distance between that next hop to the destination. A fuzzy logic system is developed and applied to the next hop selection mechanism. Thus the authors developed 27 rules for the fuzzy logic set to select the next hop node.

Braided multipath routing (BMR): Ganesan et al. [69] proposed braided multipath routing protocol. In BMR protocol each node discovers alternate best paths from a source to a destination during the route discovery process.

Truth multipath routing protocol (TMRP): Wang et al. [70] proposed TMRP that can be suitable for network with non-cooperative nodes referred as selfish nodes depending upon the resource availability at each node, the cost of forwarding a packet is measured.

Ad hoc on-demand multipath distance vector routing (AOMDV): AOMDV based on traditional AODV was proposed by Marina and Das [71]. The main objective of this protocol is to establish a multiple loop free and link-disjoint paths. The proposed metric namely "advertise hop count "is used in this protocol. The advertised hop count for a node is defined as the maximum acceptable hop count for any path recorded at the node.

Disjoint multipath routing using colored trees: Ramasubramanian et al. [72] developed a loopfree multipath routing protocol using a pair of trees that are red and blue in colors. Thus, a pair of colored trees is constructed by this process.

Scalable multipath on-demand routing (SMORT): SMORT was developed by Reddy and Raghavan [73]. The major objective of this protocol is to minimize the routing overhead occurred during route break recovery and to increase the scalability.

Split multipath routing (SMR): Lee et al., [74] proposed SMR protocol that forms and uses multiple routes of maximally disjoint paths. The overhead caused by route recovery process is minimized by establishing a multiple path from source to destination. **Table 6** illustrates the comparative analysis of multipath routing protocols.

3.6 Multicast routing protocols

In multicasting routing, the data are transmitted from one source to multiple destinations. Multicast protocols can be categorized into two types, namely tree-based multicast and mesh based multicast. The tree based multicast routing protocols utilize the network resource in efficient manner. Mesh based protocols are robust due to formation of many redundant paths between the nodes and in high packet delivery ratio.

Ad hoc multicast routing protocol (AMRoute): Xie et al. [75] developed AMRoute, with main design objective are: scalability and robustness. In ad hoc network with highly dynamic mobile nodes, the control packets overhead are high due to maintenance of multi cast tree.

Protocol	Proactive/reactive	Loops	Route metrics	Route cache
CHAMP	Reactive	Yes	Shortest path	Yes
AOMDV	Reactive	No	Advertised hop count	No
SMR	Reactive	No	Least delay	No
TMRP	Reactive	No	Auction winner	No
SMORT	Reactive	No	Shortest path	Yes
Ramasubramanian et al.	Proactive	No	Preferred neighbor	Route table

Table 6.
Comparative analysis of multipath routing protocols.

Adaptive demand-driver multicast routing (ADMR): ADMR, on-demand multicast routing algorithm, developed by Jetchera and Johnson [76]. This protocol does not support any non on-demand components. ADMR, uses a source based forwarding trees and monitors the traffic pattern and rate of the source. ADMR navigates back to the normal mode, when the mobility of the node is reduced.

Differential destination multicast (DDM): Ji and Corson [77] proposed the DDM algorithm. DDM has two important characteristics features: 1. the sender node will have full control over the members of group nodes. 2. Source node, encodes the address within each data packets header on an in-band fashion.

Dynamic core based multicast routing (DCMP): Das et al. [78] proposed DCMP source initiated multicast protocol with an objective to increase the scalability and efficiency as well as to decrease the overhead. In this protocol the source as been classified into active, core active and passive. A core active source can support up to maximum of MaxPassSize passive resource and the hop distance between them is limited by the MaxHop parameter.

Adhoc QoS multicasting (AQM): AQM protocols developed by Bur and Ersoy [79]. In this protocol QoS of the neighboring node monitored and maintained as well as used for efficient multicast routing. Node announces the QoS status during the session initiation phase to join a session, the nodes executes request-reply–reverse procedure, ensures the QoS information is updated and a possible route is chosen session is initiated by a session initiator node.

Content based multicast (CBM): CBM developed by Zhou and Singh [80]. In CBM the nodes collect information about threats and resource at a time period t and distance d away from the location of the node.

Energy efficient multicast routing: Li et al. [81] proposed an energy efficient multicast routing protocol. The authors constructed a weighted network graph by considering the transmission power of each node as a weight between edges. Each node has only information regarding their neighbors. The objective of minimum energy multicast (MEM), problem is to develop the multicast tree with a minimum total energy cost. In this approach, multicast tree is formed by nodes within the highest energy efficiency.

QoS multicast routing protocols for clustering mobile ad hoc networks (QMRPCAH): QMRPCAH, QoS aware multicast routing protocol for clustered ad hoc network was developed by Layuan and Chunlin [82]. It enhances scalability and flexibility.

Epidemic-based reliable and adaptive multicast for mobile ad hoc networks (Eramobile): Eramobile, highly reliable and an adaptive multicast protocol proposed by Ozkasap et al. [83]. In this protocol bio-inspired epidemic methods are utilized in multicast operation in order to support dynamic and topology changes due the unpredictable mobility of the nodes in the network. **Table 7** illustrates the comparative analysis of multicast routing protocol.

3.7 Location-aware routing protocols

The geographical information about a node is collected by another node by using GPS mechanism. Location-aware routing protocols are efficiently supports to improve the scalability of the ad hoc network.

Location aided routing (LAR): Ko and Vaidya [3] presented the LAR protocol, is based on directed flooding strategies. Two different LAR schemes are proposed to determine whether a node is within the request zone.

Distance routing effect algorithm (DREAM): DREAM proposed by Basagni et al. [2] utilizes location information measured using GPS system and speed information of data packet for routing. The working principal of this protocol is a part proactive and reactive in nature.

Protocol	RS	Core/ broadcast	Route metrics	Forwarding strategy	Route repository	Critical node
DCMP	F	Core	New route	Source routing	RT	No
ADMR	H	Neither	Link breaks	Flooding/tree based	RT	Yes
AMRoute	H	Core	Unicast operation	Shared tree	Based upon algorithm	Yes
Li et al.	F	Neither	Minimum energy	Source routing	RC	No
QMRPCAH	H	Broadcast	QoS	Bordercast	RT	Yes
AQM	F	Core	QoS	Source routing	RT	No
CBM	F	Core	Threat arrival	Limited broadcast	RC	Yes
DDM	F	Neither	SP	Source routing	None	No
EraMobile	F	Neither	Randomly selected	Local broadcast	None	Yes

RS = routing structure; H = hierarchical; F = flat routing repository; RC = route cache; RT = route table.

Table 7.
Comparative analysis of multicast routing protocols.

Greedy perimeter stateless routing (GPSR): GPSR algorithm by Karp and Kung [15] supports scalability and mobility. The protocol exploits greedy forwarding strategies, a node forward the packet to neighbors that is closer to the destination than itself until the destination is reached.

Dynamic route maintenance (DRM) for geographical forwarding: Chou et al. [84] developed a dynamic beacon based geographical routing algorithm. **Improvements to location-aided routing through directional count restrictions:** Colagross et al. [85] defined a scheme using the count threshold value that keeps track of number of duplicate broadcast packet received by a node. The main objective is to minimize the control packets overhead by decreasing duplicate route discovery packets.

Adaptive location aided mobile ad hoc network routing (ALARM): ALARM algorithm proposed by Boleng and Camp [86], exploits link duration as mobility feedback for adaptation and for evaluating the performance improvement, location informed are used.

A region based routing protocol for wireless mobile ad hoc networks (REGR): REGR by Liu et al. [87] proposed dynamically established a pre-routing region between source-destination pair. The two main features about this protocol are: REGR route creation and REGR route update.

Maximum expectation within transmission range (MER): Kwin and Shroff [88] presented the MER, a location- aware protocol. Each node in the location aware routing use location monitoring tool namely GPS.

SOLAR: Ghosh et al. [89] proposed a framework called ORBIT to achieve the macro-level mobility. ORBIT is defined as an orbital movement pattern of mobile users along specific places called hubs.

Geographical landmark routing (GLR): Kim [90] described GLR algorithm, GLR gives solutions to two major routing issues namely blind detouring problem and the triangular routing problem.

Secure position based routing protocol: Song et al. [91] described a highly secure geographical forwarding (SGF) algorithm. SGF provides source authentication message authentication and message integrity.

On-demand geographical path routing (OGRP): OGRP is an efficient, stateless and scalable routing protocols by Giruka and Singhal [92]. OGRP exploits the features of greedy forwarding, reactive route discovery and source routing.

Location aided knowledge extraction routing for mobile ad hoc networks (LAKER): LAKER protocol proposed by Li and Mohapatra [93]. This protocol combines the features of caching strategy and limited flooding area to decrease the network overhead. **Table 8** describes the comparative analysis of location-aware routing protocols.

3.8 Geographical multicast (Geocast) routing protocols

Geocast routing protocols have the combined features of both geographical and multicast routing protocols. The major advantage of Geocast routing protocols are performance improvement and minimizing the control overhead.

Geocasting in mobile ad hoc networks (GeoTORA): Ko and Vaidya [8] proposed the GeoTORA protocol, is based upon the unicast TORA routing protocol.

Geocast protocol for mobile ad hoc network based on GRID (GEOGRID): GeoGRID routing protocol was developed by Liao et al. [7] GeoGrid extends on the unicasting routing protocol GRID. GeoGRID exploit location information in route discovery to define the forwarding zone or geographical area.

Direction guided routing (DGR): An and Papavassilliou [94] designed DGR algorithm based on clustering mechanism. In DGR, the nodes in the network are grouped into clusters and the cluster head is elected using the techniques such a mobile clustering algorithm (MCA).

Geocast adaptive mesh environment for routing (GAMER): GAMER protocol developed by Camp and Liu [95] is based on the mobility nature of nodes. This protocol exploits the mesh creation approach. **Table 9** illustrates the Geocast routing protocols comparative analysis.

Protocol	Forwarding mechanism	Loop	Route metric	Scalability	Robustness
LAR	Directional flooding	No	Hop count	No	No
DREAM	Flooding	No	Hop count	No	No
GPSR	Greedy flooding	Yes	SP	Yes	No
Colargrosso et al.	Directional flooding	No	Hop count	No	No
ALARM	Directional flooding	Yes	Hops and mobility	Yes	No
REGR	Directional flooding	Yes	SP	Yes	No
LAKER	Directional flooding	No	Hop count	No	No
OGPR	Source routing	Yes	SP	Yes	Yes
SOLAR	Greedy geographic forwarding	No	SP	No	No
GLR	Source routing	Yes	SP	Yes	No
MER	Greedy geographic forwarding	No	Maximum expectation	No	Yes

Route Metric SP = shortest path; LSP = local shortest path; WDG = weighted distance gain; CC = communication complexity; H = high; M = medium; L = low.

Table 8.
Comparative analysis of location-aware routing protocols.

Protocol	RS	Core/broadcast	Route metrics	Forwarding strategy	Route repository	Critical node
DGR	H	Core	SP	Limited flooding	RC	Yes
GAMER	F	Core	SP	Source routing	RC	No
GeoGrid	H	Core	Hop count	Flooding or ticket based	None	No
GeoTora	H	Broadcast	SP	Limited flooding	RT	Yes

RS = routing structure; H = hierarchical; F = flat; SP = shortest path; RC = route cache; RT = route table.

Table 9.
Geocast routing protocol comparison.

3.9 Power-aware routing protocols

In ad hoc network, performance and lifetime of the nodes depends upon the power consumed by them. Thus energy efficiency is an important and challenging issue in designing power-aware routing protocols.

Device and energy aware routing (DEAR): DEAR, a power-aware protocol for heterogeneous network is proposed by Arun Avudainayagam et al. [96] The protocol is designed for heterogeneous network that consist of two different categories of nodes namely: battery powered nodes and externally powered nodes.

Routing and channel assignment for low power transmission in PCS: Scott and Bombos [97] gave a proposal for reducing the transmission power in PCS network. The author's goal is to increase the network lifetime of the individual nodes.

Energy conserving routing in wireless ad hoc networks: Chang et al. [98] states that shortest route is the routes with the least energy cost. This leads to a conclusion, more energy will be consumed by the nodes along the shortest paths, whereas the battery power of the other nodes in the network remains unused.

CLUSTERPOW and MINPOW: Kawadia and Kumar [99] developed three different power-aware algorithms namely: CLUSTERPOW, tunneled CLUSTERPOW and MINPOW. A route chosen by this protocol guarantees that each hop in the route has a maximum transmit power capacity.

Interference aware cooperative routing: Mahmood and Comanicics [100] proposed two algorithms, with a goal to maximize the throughput and minimize energy consumption. The algorithms are designed specifically to CDMA based ad hoc sensor network.

Minimum energy hierarchical dynamic source routing (MEHDSR): Tarique and Tepa [101] developed two energy-aware protocols namely MEDSR and HMEDSR based on DSR, the traditional source initiated routing protocols.

Power conserving routing with entropy-constrained algorithm: Karayiannis and Nadella [102] present a routing algorithm with an objective to reduce the overhead involved with route discovery. The authors applied the concept of entropy to develop the power-aware routing algorithm. Thus, two specific implementations are discussed.

1. Single performance metrics: optimizing the route with link cost metrics.

2. Multiple performance metrics: optimizing the route with link cost and link reliability.

This algorithm proves that the entropy constrained algorithms can improve the network lifetime.

Protocol	RT	Type	Path strategy	Routing metrics	Scalability	Robustness	Critical node
DEAR	Yes	Global	Single-path	Based upon 'device type'	No	Yes	No
Scott and Bombos	No	Centralized	Single-path	Multiple constrained SP	Yes	No	No
CLUSTERPOW	Yes	Clustered	Single-path	Total consumed power	Yes	Yes	Yes
Mahmood and Comaniciu	No	Distributed	Single-path	Energy and interference	No	No	No
MEHDSR	No	Global	Single-path	SP or next available link	Yes	No	No
Karayiannis and Nadella	No	Distributed	Single-path	Link cost and link reliability	Yes	No	No

Routing metrics: SP = shortest path.

Table 10.
Comparison of power aware routing protocols.

Table 10 illustrates the comparative analysis of power aware routing protocol comparison.

4. Conclusion

In this paper, a survey is performed on various routing algorithms including the traditional routing algorithms namely table-driven and source-initiated routing algorithms. Thus the ad hoc routing algorithm is divided into nine categories: (1) source-initiated (reactive or on-demand), (2) table-driven (proactive), (3) hybrid, (4) hierarchical, (5) multipath, (6) multicast, (7) location aware, (8) geographical multicast, (9) power-aware. Even though each protocol classes have different operational mechanism their all come under one roof by having common aim to minimize packet overhead, maximize throughput and minimize end-end delay. In this survey, the major routing issues faced by the routing protocols are discussed and effective study about the various categories of routing algorithm along with a comparative study is performed.

Author details

Alagan Ramasamy Rajeswari
Sethu Institute of Technology, Tamil Nadu, India

*Address all correspondence to: arrajeswari.2015@gmail.com

IntechOpen

References

[1] Alotaibi E, Mukherjee B. A survey on routing algorithms for wireless ad-hoc and mesh networks. Computer Networks. 2012:944-955

[2] Basagni S, Chlamtac I, Syrotiuk VR, Woodward BA. A distance routing effect algorithm for mobility (DREAM). In: Proceedings of the Fourth Annual ACM International Conference on Mobile Computing and Networking (MobiCom'98). 1998. pp. 76-84

[3] Ko Y-B, Vaidya NH. Location-aided routing in mobile ad-hoc networks. In: Proceedings of the Annual ACM International Conference on Mobile Computing and Networking (MobiCom'98). October 1998. pp. 66-75

[4] Ko Y, Vaidya NH. Geo-casting in mobile ad-hoc networks: Location-based multicast algorithms. In: Proceedings of the Second IEEE Workshop in Mobile Computing Systems and Applications (WMCSA'99); Vols. 25–26. February 1999. pp. 101-110

[5] Stojmenovic X. Lin, loop-free hybrid single-path/flooding routing algorithms with guaranteed delivery for wireless networks. IEEE Transactions on Parallel and Distributed Systems. 2001;**12**(10): 1023-1032

[6] Park V, Corson S. A highly adaptive distributed routing algorithm for mobile wireless networks. In: Proceedings of the Sixteenth Annual Joint Conference of the IEEE Computer and Communications Societies (INFOCOM'97); Vol. 3. 1997. pp. 1405-1413

[7] Liao W-H, Tseng Y-C, Lo K-L, Sheu J-P. GeoGrid: A geocasting protocol for mobile ad hoc networks based on grid. Journal of Internet Technology. 2000;**1**(2):23-32

[8] Ko Y, Vaidya NH. GeoTORA: A protocol for geo-casting in mobile ad-hoc networks. In: Proceedings of the IEEE International Conference on Network Protocols. 2000. p. 240

[9] Haas ZJ, Pearlman M. The zone routing protocol (ZRP) for ad-hoc networks. IETF Internet Draft. July 2002. Available from: draft-ietf-manet-zone-zrp-04.txt

[10] Huang Q, Lu C, Roman G-C. Mobicast: Just-in-time multicast for sensor networks under spatiotemporal constraints. Lecture Notes in Computer Science. 2003;**2634**:442-457

[11] Eppstein D. Spanning trees and spanners. Tech. Report No. ICS-TR-96-16. Irvine: University of California; 1996

[12] Rao A, Ratnasamy S, Papadimitriou C, Shenker S, Stoica I. Geographic routing without location information. In: Proceedings of the Annual ACM International Conference on Mobile Computing and Networking (MobiCom'03); September. 2003. pp. 96-108

[13] Xing G, Lu C, Pless R, Huang Q. On greedy geographic routing algorithms in sensing-covered networks. In: Proceedings of the Fifth ACM Annual International Symposium on Mobile Ad-Hoc Networking and Computing, ACM Special Interest Group on Mobility of Systems, Users, Data, and Computing (SIGMOBILE'04). 2004. pp. 31-42

[14] Leong B, Liskov B, Morris R. Geographic routing without planarization. In: USENIX Symposium on Networked Systems Design and Implementation. 2006

[15] Karp B, Kung H. GPSR: Greedy perimeter stateless routing for wireless networks. In: Proceedings of the Sixth Annual ACM International Conference on Mobile Computing and Networking (MobiCom'00). 2000. pp. 243-254

[16] Li J, Janotti J, De Coutu DSJ, Karger DR, Morris R. A scalable location service for geographic ad-hoc routing. In: Proceedings of the Sixth ACM Annual International Conference on Mobile Computing and Networking, ACM Special Interest Group on Mobility of Systems, Users, Data, and Computing (SIGMOBILE'00). 2000. pp. 120-130

[17] Eriksson J, Faloutsos M, Krishnamurthy S. Scalable ad-hoc routing: The case for dynamic addressing. In: Proceedings of the Annual Joint Conference of the IEEE Computer and Communications Societies (INFOCOM'04). 2004

[18] Camara D, Loureiro A. A novel routing algorithm for ad-hoc networks. In: Proceedings of the Thirty-third Hawaii International Conference on System Sciences. 2000

[19] Caleffi M, Ferraiuolo G, Paura L. Augmented tree-based routing protocol for scalable ad-hoc networks. In: Proceedings of the Third IEEE International Workshop on Heterogeneous Multi-Hop Wireless and Mobile Networks (MHWMN'07). 2007

[20] Lindgren A, Schelen O. Infrastructured ad-hoc networks. In: Proceedings of theInternational Conference on Parallel Processing (International Workshop on Ad-Hoc Networking (IWAHN 2002)). 2002. pp. 64-70

[21] Gomez J, Campbell AT, Naghshineh M, Bisdikian C. PARO: Supporting dynamic power-controlled routing in wireless ad-hoc networks. Proceedings of the Wireless Networks. 2003;**9**(5):443-460

[22] Djenouri D, Badache N. Dynamic source routing power-aware. International Journal of Ad-Hoc and Ubiquitous Computing (IJAHUC'06). 2006;**1**(3):126-136

[23] Singh S, Raghavendra C. PAMAS and PAMAS-power aware multi access protocol with signaling ad-hoc networks. ACM Computer Communication Review. 1998;**28**(3):526

[24] Aguayo D, Bicket J, Morris R. SrcRR: A high-throughput routing protocol for 802.11 mesh networks. 2003. Available from: http://pdos.csail. mit.edu/rtm/srcrr-draft.pdf

[25] Draves R, Padhye J, Zill B. The architecture of the link quality source routing protocol. In: Technical Report MSR-TR-2004-57; Microsoft Research. 2004

[26] Yang Y, Wang J, Kravets R. Load-balanced routing for mesh networks. ACM Mobile Computing and Communications (SIGMOBILE'06). 2006;**10**(4):3-5

[27] Yang Y, Wang J, Kravets R. Interference-Aware Load Balancing for Multihop Wireless Networks, Technical Report UIUCDCS-R-2005-2526. Department of Computer Science, University of Illinois at Urbana-Champaign; 2005

[28] Feng J, Xia R, Zhou H. Interference-aware load balanced routing in wireless mesh networks. In: Proceedings of the International Conference Wireless Communications, Networking and Mobile Computing (WiCom'07); September. 2007. pp. 1730-1734

[29] Popa L, Rostamizadeh A, Karp RM, Papadimitriou C, Stoica I. Balancing traffic load in wireless networks with curveball routing. In: Proceedings of the Eighth ACM International Symposium on Mobile and ad-hoc Networking & Computing (MobiHOC'07). 2007. pp. 170-179

[30] Johnson DB, Maltz DA, Broch J. DSR: The dynamic source routing protocol for multi-hop wireless ad hoc networks. In: Perkins CE, editor. Ad

Hoc Networking. Addison-Wesley; 2001. pp. 139-172. (Chapter 5)

[31] Perkins C, Royer E. Ad hoc on-demand distance vector routing. In: Proceedings of the Second IEEE Workshop on Mobile Computing Systems and Applications. 1999. pp. 99-100

[32] Toh C-K. Associativity-based routing for ad-hoc mobile networks. Wireless Personal Communications Journal. 1997;**4**(2 Special Issue on Mobile Networking and Computing Systems):103-139

[33] Dube R, Rais CD, Wang KY, Tripathi SK. Signal stability-based adaptive routing (SSA) for ad hoc mobile networks. IEEE Personal Communications Magazine. 1997:36-45

[34] Gunes M, Sorges U, Bouazizi I. Ara-the ant-colony based routing algorithms for manets. In: Proceedings of IEEE ICPP Workshop on Ad Hoc Networks (IWAHN); August. 2002. pp. 1599-1615

[35] Garcia-Luna-Aceves J, Mosko M, Perkins C. A new approach to on demand loop-free routing in networks using sequence numbers. Computer Networks. 2006;**50**(10):1599-1615

[36] Wang Y-H, Chao C-F. Dynamic backup routes routing protocol for mobile ad hoc networks. Information Sciences. 2006;**176**(2):161-185

[37] Xue Q, Ganz A. Ad hoc QoS on-demand routing (AQOR) in mobile ad hoc networks. Journal of Parallel Distributed Computing. 2003;**63**(2): 154-165

[38] Rosati L, Berioli M, Reali G. On ant routing algorithms in ad hoc networks with critical connectivity. Ad Hoc Networks. 2008;**6**(6):827-859

[39] Raju J, Garcia-Luna-Aceves J. A new approach to on-demand loop free

multipath routing. In: Proceedings of the Eight IEEE International Conference on Computer Communications and Networks (IC3N); April. 1999. pp. 522-527

[40] Ahn C. Gathering-based routing protocol in mobile ad hoc networks. Computer Communications. 2006; **30**(1):202-206

[41] Beraldi R, Querzoni L, Baldoni R. A hint-based probabilistic protocol for unicast communications in MANETs. Ad Hoc Networks. 2006;**4**(5):547-566

[42] Goff T, Abu-Ghazaleh NB, Phatak DS, Kahvecioglu R. Preemptive routing in ad hoc networks. Journal of Parallel Distributed Computing. 2003; **63**(2):123-140

[43] Rangarajan H, Garcia-Luna-Aceves J. Efficient use of route requests for loop-free on-demand routing in ad hoc networks. Computer Networks. 2007;**51**(6):1515-1529

[44] Perkins C, Bhagwat P. Highly dynamic destination-sequenced distance-vector routing (DSDV) for mobile computers. In: ACM SIGCOMM; August–September. 1994. pp. 234-244

[45] Clausen T, Jacquet P, Laouiti A, Muhlethaler P, Qayyum A, Viennot L. Optimized link state routing protocol for ad hoc networks. In: Proceedings of IEEE INMIC; December. 2001. pp. 62-68

[46] Villasenor-Gonzalez L, Ge Y, Lamont L. HOLSR: A hierarchical proactive routing mechanism for mobile ad hoc networks. IEEE Communications Magazine. 2005;**43**(7):118-125

[47] Munaretto A, Fonseca M. Routing and quality of service support for mobile ad hoc networks. Computer Networks. 2007;**51**(11):3142-3156

[48] Murthy S, Garcia-Luna-Aceves JJ. An efficient routing protocol forwireless networks. MONET. 1996;**1**(2):273-282

[49] Garcia-Luna-Aceves JJ, Spohn M. Source-tree routing in wireless networks. In: Proceedings of IEEE ICNP; October–November. 1999. pp. 273-282

[50] Chiang WLC, Wu H, Gerla M. Routing in clustered multihop, mobile wireless networks. In: Proceedings of IEEE SICON; April. 1997. pp. 197-211

[51] Samar P, Pearlman M, Haas S. Independent zone routing: An adaptive hybrid routing framework for ad hoc wireless networks. IEEE/ACM Transactions on Networking. 2004; **12**(4):595-608

[52] Joa-Ng M, Lu I-T. A peer-to-peer zone-based two-level link state routing for mobile ad hoc network. IEEE Journal on Selected Areas in Communications. 1999;**17**(8):1415-1425

[53] Pei G, Gerla M, Hong X. LANMAR: Landmark routing for large scale wireless ad hoc networks with group mobility. In: Proceedings of ACM MobiHoc; August. 2000. pp. 11-18

[54] Aggelou GN, Tafazolli R. RDMAR: A bandwidth-efficient routing protocol for mobile ad hoc networks. In: Proceedings of ACM WOWMOM; August. 1999. pp. 26-33

[55] Radhakrishnan S, Rao N, Racherla G, Sekharan C, Batsell S. DST —A routing protocol for ad hoc networks using distributed spanning trees. In: Proceedings of IEEE WCNC; September. 1999. pp. 100-104

[56] Nikaein N, Labiod H, Bonnet C. DDR: Distributed dynamic routing algorithm for mobile ad hoc networks. In: Proceedings of ACM MobiHoc; August. 2000. pp. 19-27

[57] Pei G, Gerla M, Chen T-W. Fisheye state routing in mobile ad hoc networks. In: Proceedings of IEEE ICDCS Workshop on Wireless Networks and Mobile Computing, April. 2000. pp. D71-D78

[58] Wang J, Osagie E, Thulasiraman P, Thulasiram R. Hopnet: A hybrid ant colony optimization routing algorithm for mobile ad hoc network. Ad Hoc Networks. 2009;**7**(4):690-705

[59] Yang C-C, Tseng L-P. Fisheye zone routing protocol: A multi-level zone routing protocol for mobile ad hoc networks. Computer Communications. 2007;**30**(2):261-268

[60] Xiaochuan X, Gang W, Keping W, Gang W, Shilou J. Link reliability based hybrid routing for tactical mobile ad hoc network. Journal of Systems Engineering and Electronics. 2008; **19**(2):259-267

[61] Bamis A, Boukerche A, Chatzigiannakis I, Nikoletseas S. A mobility aware protocol synthesis for efficient routing in ad hoc mobile networks. Computer Networks. 2008; **52**(1):130-154

[62] Sivakumar R, Sinha P, Bharghavan V. CEDAR: A core-extraction distributed ad hoc routing algorithm. IEEE Journal on Selected Areas in Communications. 1999;**17**(8): 1454-1465

[63] Iwata A, Chiang C, Pei G, Gerla M, Chen T. Scalable routing strategies for ad hoc wireless networks. IEEE Journal on Selected Areas in Communications. 1999;**17**(8):1369-1379

[64] Eriksson J, Faloutsos M, Krishnamurthy SV. Scalable ad hoc routing: The case for dynamic addressing. Proceedings of IEEE INFOCOM. 2004;**2**:1108-1119

[65] Xu K, Hong X, Gerla M. Landmark routing in ad hoc networks with mobile backbones. Journal of Parallel and Distributed Computing. 2003;**63**(2): 110-122

[66] Valera A, Seah WKG, Rao SV. Cooperative packet caching and shortest multipath routing in mobile ad hoc networks. Proceedings of IEEE INFOCOM. 2003;**1**:260-269

[67] Mavropodi R, Kotzanikolaou P, Douligeris C. SecMR—A secure multipath routing protocol for ad hoc networks. Ad Hoc Networks. 2007;**5**(1): 87-99

[68] Liang Q, Ren Q. Energy and mobility aware geographical multipath routing for wireless sensor networks. In: Proceedings of the IEEE Wireless Communications and Networking Conference, Vol. 3; March. 2005. pp. 1867-1871

[69] Ganesan D, Govindan R, Shenker S, Estrin D. Highly-resilient, energy-efficient multipath routing in wireless sensor networks. ACM SIGMOBILE Mobile Computing and Communications Review. 2001;**5**(4):11-25

[70] Wang Y, Giruka V, Singhal M. Truthful multipath routing for ad hoc networks with selfish nodes. Journal of Parallel and Distributed Computing. 2008;**68**(6):778-789

[71] Marina MK, Das S. On-demand multi path distance vector routing in ad hoc networks. In: Proceedings of IEEE ICNP. 2001. pp. 14-23

[72] Ramasubramanian S, Krishnamoorthy H, Krunz M. Disjoint multipath routing using colored trees. Computer Networks. 2007;**51**(8): 2163-2180

[73] Reddy LR, Raghavan S. SMORT: Scalable multipath on-demand routing for mobile ad hoc networks. Ad Hoc Networks. 2007;**5**(2):162-188

[74] Lee S, Gerla M. Split multipath routing with maximally disjoint paths in ad hoc networks. In: Proceedings of the IEEE ICC. 2001. pp. 3201-3205

[75] Xie J, Talpade RR, McAuley A, Liu M. Amroute: Ad hoc multicast routing protocol. MONET. 2002;**7**(6): 429-439

[76] Jetcheva J, Johnson D. Adaptive demand-driven multicast routing in multi-hop wireless ad hoc networks. In: Proceedings of ACM MobiHoc. October 2001. pp. 33-44

[77] Ji L, Corson M. Differential destination multicast-a manet multicast routing protocol for small groups. Proceedings of IEEE INFOCOM. 2001;**2**: 1192-1201

[78] Das SK, Manoj BS, Murthy CSR. A dynamic core based multicast routing protocol for ad hoc wireless networks. In: Proceedings of ACM MobiHoc. June 2002. pp. 24-35

[79] Bür K, Ersoy C. Ad hoc quality of service multicast routing. Computer Communications. 2005;**29**(1):136-148

[80] Zhou H, Singh S. Content based multicast (CBM) in ad hoc networks. In: Proceedings of ACM MobiHoc. August 2000. pp. 51-60

[81] Li D, Liu Q, Hu X, Jia X. Energy efficient multicast routing in ad hoc wireless networks. Computer Communications. 2007;**30**(18):3746-3756

[82] Layuan L, Chunlin L. A QoS multicast routing protocol for clustering mobile ad hoc networks. Computer Communications. 2007;**30**(7):1641-1654

[83] Ozkasap O, Genc Z, Atsan E. Epidemic-based reliable and adaptive multicast for mobile ad hoc networks. Computer Networks. 2009;**53**(9): 1409-1430

[84] Chou C-H, Ssu K-F, Jiau H. Dynamic route maintenance for geographic forwarding in mobile ad hoc networks. Computer Networks. 2008; **52**(2):418-431

[85] Colagrosso M, Enochs N, Camp T. Improvements to location-aided routing through directional count restrictions. In: Proceedings of International Conference on Wireless Networks (ICWN). June 2004. pp. 924-929

[86] Boleng J, Camp T. Adaptive location aided mobile ad hoc network routing. In: Proceedings of IEEE International Performance, Computing, and Communications Conference (IPCCC). April 2004. pp. 423-432

[87] Liu Y, Hu X, Lee M, Saadawi T. A region-based routing protocol for wireless mobile ad hoc networks. IEEE Network. 2004;**18**(4):12-17

[88] Kwon S, Shroff N. Geographic routing in the presence of location errors. Computer Networks. 2006;**50**(15):2902-2917

[89] Ghosh J, Philip S, Qiao C. Sociological orbit aware location approximation and routing (solar) in manet. Ad Hoc Networks. 2007;**5**(2):189-209

[90] Na J, Kim C-K. Glr: A novel geographic routing scheme for large wireless ad hoc networks. Computer Networks. 2006;**50**(17):3434-3448

[91] Song J-H, Wong V, Leung V. Secure position-based routing protocol for mobile ad hoc networks. Ad Hoc Networks. 2007;**5**(1):76-86

[92] Giruka V, Singhal M. A self-healing on-demand geographic path routing protocol for mobile ad-hoc networks. Ad Hoc Networks. 2007;**5**(7):1113-1128

[93] Li J, Mohapatra P. LAKER: Location aided knowledge extraction routing for mobile ad hoc networks. In: Proceedings of IEEE WCNC; March. 2003. pp. 1180-1184

[94] An B, Papavassiliou S. Geomulticast: Architectures and protocols for mobile ad hoc wireless networks. Journal of Parallel Distributed Computing. 2003; **63**(2):182-195

[95] Camp T, Liu Y. An adaptive mesh-based protocol for geocast routing. Journal of Parallel Distributed Computing. 2003;**63**(2):196-213

[96] Avudainayagam A, Lou W, Fang Y. Dear: A device and energy aware routing protocol for heterogeneous ad hoc networks. Journal of Parallel Distributed Computing. 2003;**63**(2):228-236

[97] Scott K, Bombos N. Routing and channel assignment for low power transmission in PCS. In: Proceedings of IEEE ICUPC, September–October. 1996. pp. 498-502

[98] Chang J-H, Tassiulas L. Energy conserving routing in wireless adhoc networks. Proceedings of IEEE INFOCOM. 2000:22-31

[99] Kawadia V, Kumar P. Power control and clustering in ad hoc networks. Proceedings of IEEE INFOCOM. 2003;**1**:459-469

[100] Mahmood H, Comaniciu C. Interference aware cooperative routing for wireless ad hoc networks. Ad Hoc Networks. 2009;**7**(1):248-263

[101] Tarique M, Tepe K. Minimum energy hierarchical dynamic source routing for mobile ad hoc networks. Ad Hoc Networks. 2009;**7**(6):1125-1135

[102] Karayiannis N, Nadella S. Power-conserving routing of ad hoc mobile wireless networks based on entropy-constrained algorithms. Ad Hoc Networks. 2006;**4**(1):24-35

Edited by Pinaki Mitra

In recent years there has been many developments in communication technology. This has greatly enhanced the computing power of small handheld resource-constrained mobile devices. Different generations of communication technology have evolved. This had led to new research for communication of large volumes of data in different transmission media and the design of different communication protocols. Another direction of research concerns the secure and error-free communication between the sender and receiver despite the risk of the presence of an eavesdropper. For the communication requirement of a huge amount of multimedia streaming data, a lot of research has been carried out in the design of proper overlay networks. The book addresses new research techniques that have evolved to handle these challenges.

IntechOpen