TAMPEREEN TEKNILLINEN YLIOPISTO
TAMPERE UNIVERSITY OF TECHNOLOGY

Mikhail Komarov

Network Challenges of Novel Sources of Big Data

Julkaisu 1444 • Publication 1444

Tampere 2016

Tampereen teknillinen yliopisto. Julkaisu 1444
Tampere University of Technology. Publication 1444

Mikhail Komarov

Network Challenges of Novel Sources of Big Data

Thesis for the degree of Doctor of Science in Technology to be presented with due permission for public examination and criticism in Tietotalo Building, Auditorium TB220, at Tampere University of Technology, on the 28th of November 2016, at 12 noon.

Tampereen teknillinen yliopisto - Tampere University of Technology
Tampere 2016

Supervisor:

Yevgeni Koucheryavy, Ph.D., Associate Professor
Department of Electronics and Communications Engineering
Tampere University of Technology
Tampere, Finland

Instructor:

Dmitri Moltchanov, Ph.D., Senior Researcher
Department of Electronics and Communications Engineering
Tampere University of Technology
Tampere, Finland

Pre-examiners:

Tapani Ristaniemi, Ph.D., Professor
Department of Mathematical Information Technology
University of Jyväskylä
Jyväskylä, Finland

Albert Lysko, Ph.D., Principal Researcher
The Council for Scientific and Industrial Research
Brummeria, Pretoria, South Africa

Opponent:

Veselin Rakocevic, Ph.D., Senior Lecturer
School of Mathematics, Computer Science & Engineering
Department of Electrical and Electronic Engineering
City University of London
London, UK

ISBN 978-952-15-3865-0 (printed)
ISBN 978-952-15-3880-3 (PDF)
ISSN 1459-2045

ABSTRACT

Networks and networking technologies are the key components of Big Data systems. Modern and future wireless sensor networks (WSN) act as one of the major sources of data for Big Data systems. Wireless networking technologies allow to offload the traffic generated by WSNs to the Internet access points for further delivery to the cloud storage systems. In this thesis we concentrate on the detailed analysis of the following two networking aspects of future Big Data systems: (i) efficient data collection algorithms in WSNs and (ii) wireless data delivery to the Internet access points.

The performance evaluation and optimization models developed in the thesis are based on the application of probability theory, theory of stochastic processes, Markov chain theory, stochastic and integral geometries and the queuing theory.

The introductory part discusses major components of Big Data systems, identify networking aspects as the subject of interest and formulates the tasks for the thesis. Further, different challenges of Big Data systems are presented in detail with several competitive architectures highlighted. After that, we proceed investigating data collection approaches in modern and future WSNs. We back up the possibility of using the proposed techniques by providing the associated performance evaluation results. We also pay attention to the process of collected data delivery to the Internet backbone access point, and demonstrate that the capacity of conventional cellular systems may not be sufficient for a set of WSN applications including both video monitoring at macro-scale and sensor data delivery from the nano/micro scales. Seeking for a wireless technology for data offloading from WSNs, we study millimeter and terahertz bands. We show there that the interference structure and signal propagation are fundamentally different due to the required use of highly directional antennas, human blocking and molecular absorption. Finally, to characterize the process of collected data transmission from a number of WSNs over the millimeter wave or terahertz backhauls we formulate and solve a queuing system with multiple auto correlated inputs and the service distribution corresponding to the transmission time over a wireless channel with hybrid automatic repeat request mechanism taken into account.

Preface

The research work that makes up this thesis has been carried out at the Department of Electronics and Communications Engineering of Tampere University of Technology (Finland) over the years 2015-2016. This manuscript aggregates the effort and vision of not just the author, but also other relevant people: colleagues, reviewers, peers, who are gratefully acknowledged here, without the intention of forgetting anyone.

First and foremost, I have been extremely fortunate to work under the supervision of Prof. Yevgeni Koucheryavy, who has significantly improved my research capabilities. I would like to thank him for his everyday support and guidance through the research process. Also, I would like to express my deepest appreciation to Dr. Dmitri Moltchanov from Tampere University of Technology. As an instructor, he helped me a lot with development of my research ideas. Without his insight, experience, and intuition it would not be possible to accomplish the research presented in this thesis.

I would like to extend my appriciation to Elina Orava for the responsiveness, promt assistance with practical matters and support.

I also would like to thank all my colleagues and friends for their everyday support on my way during all the processes connected with conducting this research.

Finally and the most importantly, I would like to express my earnest gratitude to my parents Mikhail and Yulia. They have taught me important and good things in life, especially, the value of education.

Mikhail Komarov, November, 2016, Tampere, Finland

Contents

List of Publications

This thesis is mainly based on the following publications:

[P1] M. Komarov "Network challenges of new sources of big data" , *In Proc. of the 17th IEEE Conference on Business Informatics (CBI 2015)*, 2015. Ch. 1. P. 27-36.

[P2] M. Komarov, D. Moltchanov "System design and analysis of UAV-assisted BLE wireless sensor systems" in *Wired/Wireless Internet Communications, Lecture Notes in Computer Science*, pp. 284–296, Springer International Publishing, 2016.

[P3] M. Komarov, B. Deng, V. Petrov, D. Moltchanov "Performance analysis of simultaneous communications in bacterial nanonetworks" in *Nano Communication Networks*, vol.8, pp.55-67, 2016.

[P4] A. Nguyen, M. Komarov, D. Moltchanov, "Coverage and Network Requirements of a Flash Crowd Monitoring System Using Users' Devices" in *Lecture Notes in Computer Science,* Springer International Publishing, 2016

[P5] V. Petrov, M. Komarov, D. Moltchanov, J. M. Jornet, and Y. Koucheryavy, "Interference Analysis of EHF/THF Communications Systems with Blocking and Directional Antennas", *In Proc. of the 2016 IEEE Global Communications Conference (GLOBECOM)*, 2016.

[P6] V. Petrov, M. Komarov, D. Moltchanov, J. M. Jornet, and Y. Koucheryavy, "Interference and SINR in Millimeter Wave and Terahertz Communication Systems with Blocking and Directional Antennas", *Accepted to IEEE Transactions on Wireless Communications*, 2016.

A detailed description of all publications is available in the second part of the thesis in Section "Summary of Publications" on page 47. Author's contributions are summarized in Section "Author's Contribution" on page 49.

List of Abbreviations

3GPP	Third Generation (3G) Partnership Project
ACK	Acknowledgment message or packet
AP	Access Point
ARQ	Automatic repeat request
BER	Bit error rate
BLE	Bluetooth Low Energy
CCSS	Cloud Computing Storage Systems
CDF	Cumulative Distribution Function
D-BMAP	Discrete-time Batch Markovian Arrival Process
DNA	Deoxyribonucleic acid
FEC	Forward Error Correction
FPT	First-passage time
GSM	Global System for Mobile (Communications)
HARQ	Hybrid Automatic repeat request
HDFS	Hadoop Distributed File Systems
IaaS	Infrastructure as a Service
IDC	International Data Corporation
IEEE	Institute of Electrical and Electronics Engineers
IETF	Internet Engineering Task Force
iid	independent identically distributed
ITU	International Telecommunications Union
IoT	Internet of Things
IP	Internet Protocol
IPT	Inter-passage time
ISM	Industrial, scientific and medical license-exempt bands
LTE	Long Term Evolution
LTE-A	LTE-Advanced
LoS	Line-of-Sight
M2M	Machine-to-Machine

MIMO	Multiple-Input and Multiple-Output
NACF	Normalized autocorrelation function
NTRS	Network Traffic Recording System
PaaS	Platform as a Service
PF	Probability function
PDU	Protocol Data Units
PPP	Poisson Point Process
QoS	Quality of Service
RAT	Radio Access Technology
RV	Random Variable
Rx	Receiver
SaaS	Software as a Service
SIR	Signal to Interference Ratio
SINR	Signal to Interference-plus-Noise Ratio
THz	Terahertz
Tx	Transmitter
UILS	User Interaction and Learning System
UAV	Unmanned Aerial Vehicle
WSN	Wireless Sensor Networks

List of Figures

Chapter 1

Introduction

1.1 Research Motivation

Over the last decade, various technologies have moved to the qualitatively new phase of development. Mobile wireless communications, new data collection, new data retrieval and new data processing techniques have been introduced together with a cloud approach to storing data. All of that with the *Internet of Everything* paradigm [2] formed a new concept of the *Big Data* [3].

Although many businesses are nowadays implementing their own data collection techniques and data analytics for their purposes, efficiency of those methods can be far from optimal not just because of the cost of data but also due to the non real-time nature of solutions, their low energy efficiency, loss of data during the data delivery process and insufficient utilization of novel data sources. The efforts in this direction are currently oriented towards unification and standardization of various processes related to Big Data models. Particularly, one of the major research questions addressed is development of efficient data analysis algorithms for getting additional value from the large arrays of heterogeneous information.

Historically, a wireless sensor network (WSN) has been considered one of the most important sources of Big Data. Importantly, it is the collection of WSNs that are conventionally treated as the sources of Big Data, not particular systems in isolation. With the advances in telecommunications technologies, the ever increasing needs from different authorities for more data, availability of various sensors on the open market, and further miniaturization of end systems, WSNs are finding new applications not only at macro-scales but on micro- and nano-scales. Furthermore, conventional WSNs continue to evolve further, covering more application scenarios. In the near future, we may witness the appearance of WSNs that would generate much more data than before and could prove to be a single source for a Big Data system.

In the above mentioned context we can identify two problems of future WSNs acting as sources for Big Data systems [P1]. These are: (i) specifying efficient data collection algorithms and (ii) solving the problem of data delivery to the Internet access points (and, eventually, to cloud storage). The first problem is related to changing the data collection algorithms in conventional WSNs such that their application scope can be effectively expanded while reducing the cost

of data and extending the lifetime of networks. For completely novel systems including those of macro- and micro-scale, we need to develop novel data collection techniques.

The second problem is related to finding a way to efficiently offload data from the networks acting as sources for Big Data systems. Indeed, with the fast growing number of systems using Big Data we understand the limitations in existing infrastructure for data collection and delivery. Although we have different wireless communication technologies available today such as Wi-Fi, LTE, etc., their characteristics may not satisfy the growing demands in terms of capacity and may not guarantee reliable delivery of the data from the source to the data collection hub. One of the potential options is to move higher in the frequency band from microwave systems (conventional Wi-Fi and LTE technologies) to millimeter wave (mm-wave, 30-300GHz) and terahertz (THz, 0.3-3THz) systems [4].

This thesis focuses on addressing the above mentioned two problems for modern and future WSNs having the potential to become new data sources for Big Data systems.

1.2 Background of Big Data and challenges

Different technological advancements of the previous decade have led to the dramatic increase in the amount of data collected from different sources and persisted in data storage. Good examples of data storage progression over time: e-mails, then to social networks, and further to data collection from various monitoring systems and sensors. Data was collected not just for the purposes of users but also for future business development. Analysis of various customer data allows companies to later develop new products adjusted accordingly to the customer. Due to the huge amount of data which needed to be collected, transferred, processed and analyzed, researchers formed key measures and characteristics of Big Data [5, 6] which are often described as “five Vs” [P1]:

- Volume. The first inherent characteristic is the amount of data presumed in Big Data applications. While this is not the only aspect, it is usually the size of the data that determines the value and potential of the data under consideration. Huge size of data places obvious constraints for data storage at the processing points and data delivery via the network.

- Variety. This is another inherent characteristic of Big Data application describing the different nature of data and also different semantics in data representation. This information is often available having the knowledge of the data sources and needs to be supplied to a pre-processing decision entity to their advantage and thus upholding the importance of Big Data.

- Velocity. The term ”velocity” in the context of Big Data refers to the speed of data generation. High rate of data generation places challenges on network delivery and data processing. One straightforward example is sensor network where the data generation process might be controlled by prescribed length of measurement intervals.

- Veracity. This term describes the ”quality” of obtained data. The quality is often understood in context of reliability and trust as the generated

data could vary greatly with respect to these metrics.

- Value. This characteristic refers to the complexity of using Big Data application for getting the results. Data management can become a very complex process, especially, when large volumes of data come from multiple sources. These data need to be linked and correlated in order to be able to grasp the information that is supposed to be conveyed by these data.

All these characteristics also generate restrictions and requirements to the networks infrastructure and management of resources. It all requires a plenty of computational resources and the analysis is often performed using computational clouds.

Cloud computing is an extremely successful paradigm of service-oriented computing, and has revolutionized the way computing infrastructure is abstracted and used. Three most popular cloud paradigms include: Infrastructure as a Service (IaaS), Platform as a Service (PaaS), and Software as a Service (SaaS). The concept however can also be extended to Database as a Service or Storage as a Service. Authors in [7] have reported about security issues associated with Big Data in cloud computing. Cloud computing plays a very vital role in protecting data, applications and the related infrastructure with the help of policies, technologies, controls, and Big Data tools. Cloud computing comes with numerous security issues because it encompasses many technologies including networks, databases, operating systems, virtualization, resource scheduling, transaction management, load balancing, concurrency control and memory management. Hence, security issues of these systems and technologies are applicable to cloud computing. Among others the challenges of security in cloud computing environments also include network level. Some research projects like in [8] conducted to the area of Big Data processing in a cloud environment have shown some dynamic load balancing methods and data stream distribution for better processing the data as these two problems are very important in terms of limited processing capabilities. In [9] there were proposed some challenges in terms of Big Data applications. Big Data introduced challenges driven by both user requirements (closeness/availability of data, real-time data and results delivery) and technical challenges (parallel computations might not be simple due to different types of data context, etc.). It brings two fundamental problems: how to deliver the data to/from the user in real time and how to properly manage computational tasks in a cloud by taking into account the inherent properties of Big Data.

1.3 Networks as a sources of Big Data

Recent advances in wireless communications and electronics have enabled the development of low-cost, low-power, multifunctional sensor nodes that are small in size and communicate untethered in a short distances. These tiny and generally simple sensor nodes consist of sensing units, data processing, and communication components [10, 11, 12]. They can be deployed to provide in-situ, real-time data about the state of the environment or different objects. A large number of such nodes deployed over large areas can potentially collaborate with each other.

1.3.1 Networks as generators of data

Even though wireless sensor networks have come a long way, a number of issues are still open and deserve further investigation. Clearly, the large amount of sensors involved can produce a lot of data which should be converted into meaningful information (analyzed). To achieve this, one needs to answer a number of questions such as what needs to be sensed, who should sense, whom the data must be passed on to, how are the data routed to the destination etc. On top of these, to answer these questions, one needs to take into account several constraints like the limited power of each sensor node, its low processing capabilities and bandwidth, the dynamic nature of the sensor field (nodes may move or "die" due to energy depletion).

Over the last decade researchers addressed this problem identifying a number of feasible solutions such as multi-path routing, clusterization, data aggregation, in-network data processing, etc. However, none of those are general enough to be applicable to an arbitrary deployment [13]. On top of this, the routed principle of WSNs adds to this problem. Indeed, networking mechanisms such as neighbor discovery, connectivity and topology maintenance, routing and packet forwarding require substantial amount of energy [14].

Quite interesting solutions for sensor networks nowadays are monitoring systems with the use of audio and video sensors while operating in the crowds of people. Though amount of data collected from such networks is extremely huge there are several open issues for researchers about data collection from such networks and data delivery to the data storages which involve energy-efficiency and network lifetime [15, 16].

An inherent feature of modern and forthcoming Big Data applications is appearance of new sources of data. In addition to classic wireless sensor network, currently, there is a great interest in micro- and nano-scale networks including bacteria-based systems [17]. This concept is an extension of macro-scale networks to extremely small devices. Applications where those networks are utilized may potentially generate huge arrays of data and can be characterized by all five Vs we introduced above.

1.3.2 Big Data delivery

The Figure 1.3.1 shows the schematic diagram of the network structure of data collection from different networks and data delivery to the data storage center where data analytics routines should apply. Wireless connection between the data delivery network and the data storage is presented there as one of the possible approaches or those case when wired connectivity does not exist and/or cannot be used due to economic reasons.

One of the critical questions considering networks as potential sources of Big Data is whether to process data in-network or transfer it to the data storage. While for the sensor networks that question also influenced on the energy-efficiency and resources distribution [18, 19], similar questions are in focus for Big Data systems. The effect of network size on the number of bytes transferred in the network during query execution is a linear growth for in-network data processing [19]. It requires more resources and takes more time to process the data and life-time of the network is reducing dramatically. In the case of the central data processing at the data storage, doubling the number of nodes

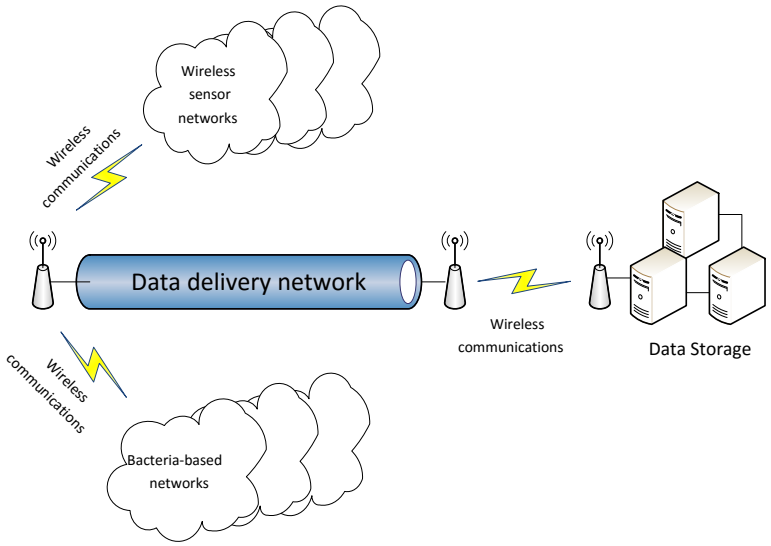


Figure 1.3.1: Network structure of data collection and delivery to the data storage.

results in 2,2 times increase of the size of transferred data. This trend holds for all cache sizes and the very same trend is also reflected in the number of messages that are passed through the network during query execution [19]. Experimental results in [20, 19] have shown that centralized data processing is a viable alternative to in-network computing under various circumstances especially when the result accuracy, the response time, and the data reusability are of primary concern, which are key parameters for Big Data systems. The key research question there is which data delivery networks should be used for Big Data systems, which would have better performance than traditional wireless communication technologies. Several challenges, which will be further analyzed in this thesis: efficient data collection from different networks including novel data sources such as nano- and bio-networks and delivery of collected data to the data storage.

1.4 Scope of the Thesis

In the thesis, we concentrate on conventional and future WSNs that are expected to act as individual sources of Big Data. We target two critical problems: (i) data collection in WSNs and (ii) collected data delivery to the Internet backbone. Accordingly, the thesis is logically divided into two corresponding parts.

We start by analyzing the scope of current challenges for Big Data and further focus on particular network challenges related to different WSNs. We analyze performance of novel bacteria-based networks and modern environmental and audio/video monitoring systems. The presented analyses are related to

efficient data collection for further delivery and processing.

We then proceed concentrating on the process of collected data delivery to the Internet backbone access point. We demonstrated that the capacity of conventional cellular system can be insufficient for a set of WSN applications including both video monitoring at macro-scale and sensor data delivery from the nano/micro scales. Moving up in the frequency band to millimeter and THz bands, we show that these systems potentially provide enough capacity to timely transfer large amounts of data.

Studying the forthcoming wireless communications systems operating in mm-wave and THz bands we will demonstrate that the interference structure in these systems is fundamentally different due to the required use of highly directional antennas and the effects of human blocking and molecular absorption. Particularly, we will show that under certain conditions these systems may operate in noise-limited regime relaxing requirements on interference mitigation techniques.

Finally, to characterize the process of collected data transmission from a number of networks over the millimeter wave or THz backhauls we will formulate and solve a queuing system with multiple auto correlated inputs and the service distribution corresponding to the transmission time over a wireless channel with hybrid automatic repeat request (HARQ) mechanism taken into account. The main conclusion is that even severe autocorrelations in individual arrival flows are of secondary importance compared to the distribution of the number of arriving packets allowing to use a simple queuing systems with uncorrelated inputs for performance analysis and dimensioning of the abovementioned backhaul technologies.

1.5 Thesis Outline

This thesis is compound, and consists of an introductory part with *five* chapters and conclusions, along with a compilation of *six* main publications referred to as [P1]-[P6].

In Chapter 2 we start with the basic description of Big Data. We discuss importance of Big Data, different models and challenges. We also briefly address those critical aspects of Big Data concept that are not targeted in the thesis. In Chapter 3 we concentrate our attention on WSNs acting as potential sources of Big Data. We analyze data collection techniques in three different WSNs including bacterial systems, conventional environment monitoring WSNs and new flash crowd monitoring systems. In Chapter 4 we consider the question of data delivery to the Internet backbone using the wireless backhauls operating in THz or mm-wave frequency bands. Chapter 5 concludes the introductory part, connecting all of the research.

The compilation part of the thesis summarizes the publications presented in this thesis and highlights the author's contribution to them.

Chapter 2

Big Data

Conceptually, the network challenges studied in this thesis are due to Big Data systems evolution. Even though according to the Gartner, Big Data is already in place with a set of associated technologies such as Internet of Things, cloud technologies and machine learning [21], we are interested in Big Data mainly in context of the networking technologies supporting data collection and data delivery. In this chapter, we review the basics of Big Data influencing the application development and networks functionality. Examples of Big Data systems complete architectures are also provided together with the associated challenges.

2.1 The importance and challenges of Big Data

Big Data is a broad term used to describe data sets so large or complex that traditional data processing applications are inadequate. There are numerous challenges when handling these data including analysis, capture, managing, search, sharing, storage, transfer, visualization, and information privacy. The term often refers simply to the use of predictive analytics or other certain advanced methods to extract value from data, and seldom to a particular size of data set [5]. According to IDC forecasts (Worldwide Semiannual Big Data and Analytics Spending Guide) published in 2016, worldwide revenues for big data and business analytics will grow from nearly $122 billion in 2015 to more than $187 billion in 2019. Typical characteristics of Big Data (or the five Vs) were presented in the Chapter 1.

The growth of Big Data is nowadays accompanied by substantial rise of the Internet of Things (IoT). With the modern rise of interest in the IoT, researchers tend to give their own definitions of this term. While the exact definition of the IoT is rather ambiguous we refer to the IoT as a way to facilitate the connection between application and services of the virtual world and the physical world of things, so that we will be able to control and sense our environment in better and more efficient way. Combining the IoT and Big Data can even facilitate new capabilities. For instance, systems configured with enough intelligence can offer "mass customization", which uses computer-aided manufacturing methods to enable individually customized products produced with efficiency and costs approaching mass production. Customers can thus have access to a new variety

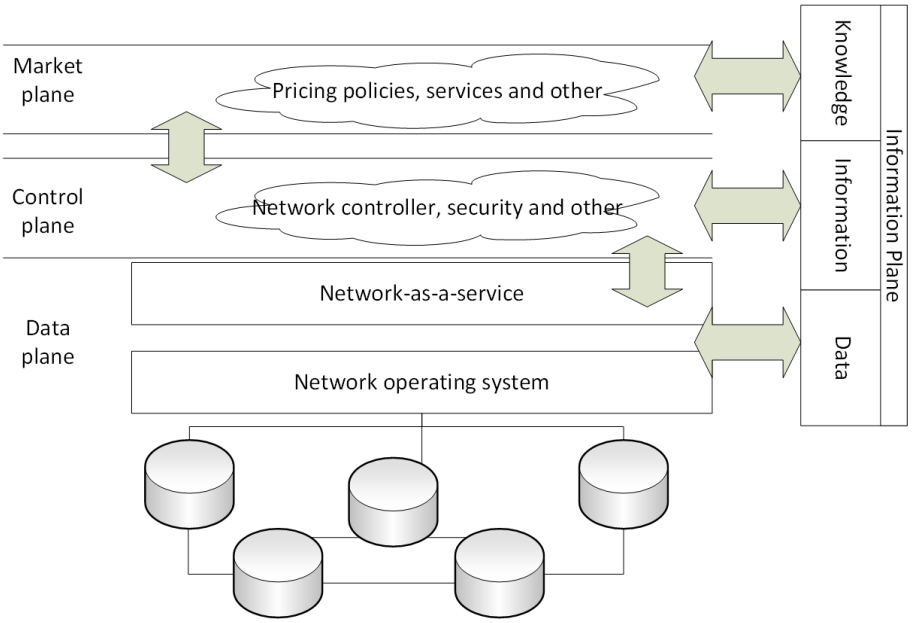


Figure 2.1.1: Data driven architecture with four planes: data, control, information, and market.

of customized products at a less than a full-custom price, while manufacturers are able to tap new markets and realize increased profits.

In the above mentioned context, it is important to remember how much data all these 'smart' things generate and transmit over the network using such technologies like sensor networks, Wi-Fi, Bluetooth etc. Big Data transform the design philosophy of the networks and complex systems as well as the Internet itself. There are usually challenges in terms of data processing time, data transmission time and delivery of the information and what is more important knowledge out of the information which should be gained out of the data collected. In [22], there were presented new data-driven architectures based on new network technologies providing data for the future Internet. Four planes are based on data and management of resources required for data collection, transmission and processing (Figure 2.1.1). Here, we present it in a simplified form to highlight the importance of data plane, which is the basis for information plane influencing on business strategies and new services and products development. As one may observe, all the services and Big Data applications depend on the network lifetime and resources [23, 9].

When Big Data are distilled and analyzed in combination with traditional enterprise data, companies can develop a more thorough and insightful understanding of their business, which can lead to enhanced productivity, a stronger competitive position and greater innovation.

Continuing the overview of the importance of Big Data, it is necessary to highlight general stages which lead to the efficient data analytics while using the IoT. While the speed is important in harvesting data from the IoT, it is important to point out that determinism is the real goal. Determinism implies that each action can reliably predict another action. The more information we

have, the more deterministic we can become since we can correlate more inputs to derive more accurate output predictions. More accurate predictions will in turn improve prognostics and health management (for instance) and other predictive technologies. This ever-improving automated sense/predict cycle leads to increased productivity due to reduced manual data collection and interaction. Better quality products are made closer to specifications and tolerances, with lower reject rates and higher throughput.

While the focus of many researchers in the past was on Big Data challenges from the top-level prospective (data visualization, data analytics and machine learning, data storages, etc., [24, 25, 26, 27, 28, 29, 30]) data collection and data delivery networks are of special importance in context of Big Data [31, 32, 23, 1, 22, 9, 33, 34].

2.2 Big Data collection and communications challenges

All the benefits of Big Data systems cannot be realized until data are first collected, which starts at the level of devices in data collection networks. Data collection techniques depend on type of devices (sensors) we are using and type of networks. There are many ways to collect data, and these data must then be gathered somewhere in order to deliver it further to the data storage. The data also could be processed on device and the processing results will be sent to the data storage, which would significantly decrease amount of data transferred through the data collection and data delivery networks. Decrease in amount of data also leads to the decrease in energy usage [15] and that is why system developers should decide whether they collect all the data at the data storage or they process some data on devices. The latter depends on devices, which generate data in the data collection networks. As a result of the concentration, we can forecast the network load and further understand which communication technologies should be used for data delivery to the data storage.

Next stage after the data delivery is management of the received data in order to store them properly. The activity of dealing with these data typically involves the use of database software. Dealing with the data is usually accomplished at the server level, since the database management is an advanced software function that requires equally capable hardware.

To a great extent the data collection, delivery, and dealing with the data activities occur silently and are not seen for the end users after the initial configuration. These activities must be in place to act as a foundation for visualizing and analyzing information. However, simply presenting users with large tables of values from a historical database will not help most people to properly understand and/or interpret the data. The reason is that most humans are visually-oriented towards graphical representations of data such as charts, graphs and other symbols.

While talking about analytics we consider that the modern computer technologies, like Hadoop Distributed File Systems (HDFS) and public cloud, can help alleviate the cardinality problem in Big Data analytics. They can be integrated to build a large and flexible network with a storage infrastructure that can change adaptively based on the need of Big Data processing requirements.

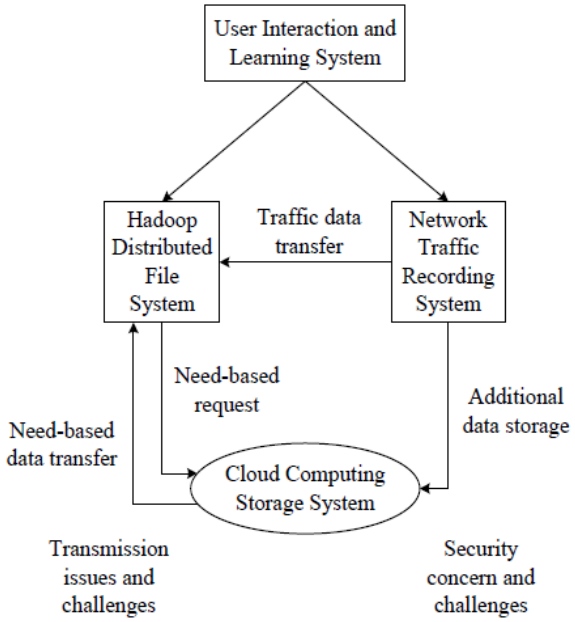


Figure 2.2.1: Network model for Big Data analytics [1].

However, this integrated model will bring several challenges that must be handled efficiently. One of the potential models, proposed in [1], is illustrated in Figure 2.2.1. It consists of four units: user Interaction and Learning System (UILS), Network Traffic Recording System (NTRS), HDFS and Cloud Computing Storage System (CCSS). The NTRS unit helps to capture network traffic and streams the traffic data to HDFS unit or CCSS unit in real-time based on the need of an additional storage. The HDFS system will also use database to store the data. The UILS unit can learn and control the additional storage and data requirements.

As we can see from the proposed model the communication cost is the major concern. The challenge here is to minimize that communication cost while satisfying the additional storage and data requirement from public cloud for processing Big Data. With reference to Figure 2.1.1, the data plane is most important one. Hence, the data collection bandwidth and latency at the data collection network are the two important network features that will affect the communication between data collection network and data delivery network. Further, the same characteristics will be important for data delivery network from the data collection networks to the data storage (cloud server). These problems and associated challenges to find solutions will affect the timing requirements of Big Data processing at HDFS and UILS.

2.3 Big Data trade-offs

As far as research is concerned, the challenges of Big Data systems are not only related to the way data are analyzed at the processing centers but to the communications and networking technologies as well. For example, should Big Data system require additional visualization or additional analytics, the networks will be used to collect additional data and deliver them using the data delivery network. If the data can be processed at the data collection network, it could potentially reduce the amount of data going through the data delivery network, thus, reducing contention and improving throughput and energy efficiency of the system itself [15, 1, 8]. The efficient data collection would influence on the performance of Big Data system as a whole. Either way, the architectures reviewed here are scalable enough to accommodate nearly any of the future network assistance models, and thus, provides a solid grounds for further evolution of Big Data systems.

In the next chapter, we focus on the performance analysis of the novel network technologies used as data collection networks and discuss their key trade-offs.

Chapter 3

Big Data collection networks

The efficiency of the data collection process in WSNs is one of the critical factors affecting the widespread adoption of such systems. For example, considering conventional macro-scale 'routed' WSNs the limited lifetime of end systems prevents their use in many promising applications affecting the cost of the obtained data. Going at micro- and nano-scales and considering bio-inspired systems such as those using bacteria for transferring data, the data collection is affected by many parameters that do not exist in macro-scale systems. Finally, there is a set of emerging applications for WSNs that require a completely different philosophy in the data collection process and may have extraordinary requirements for the access network rates. One of the examples of such systems we consider in this chapter is a flash-crowd monitoring system.

From the network perspective, we are interested in system capacity; from the user's perspective, we mostly care about throughput and power efficiency. Since these questions are difficult to address analytically for all the systems, we sometimes need to resort to system level simulations to see the trade-offs between the metrics of interest. In this chapter, we provide several key design details that make the implementation of Big Data collection networks a reality.

Performance evaluation is an extremely complex subject. Essentially, with the right environment assumptions, almost any communications technology can be shown to be exceptionally good or hopelessly bad depending on the objective of the evaluator. While we cannot claim the unbiased view, we at least aim at transparency in the evaluation. In this chapter, for each considered system, we discuss design details, and then perform the analysis of the data collection process using either system level simulations or, whenever feasible, analytical performance analysis.

3.1 Conventional wireless sensor networks

The concept of conventional routed WSNs and the associated set of applications are known for years. There are a number of communications technologies standardized and several vendors are offering the complete solutions on the market. Still, the use of these systems is limited to specific often life-saving ap-

plications such as forest fire monitoring, earthquake monitoring, tsunami monitoring, etc. [35, 36, 37, 38, 39]. At the same time, there are many applied fields, where the use of WSNs would be of high practical demand but prevented by the high cost of maintenance, e.g., agriculture fields, mines monitoring systems. In this subsection, we introduce an example of conventional WSN and identify the routed nature as the most critical factor for high energy consumption and associated maintenance cost. We then proceed introducing a new automated data collection approach allowing to significantly extend the battery life of end systems prolonging the lifetime of the network. The detailed description could be found in [P2].

3.1.1 System model

Uneven energy consumption is one of the reasons for limited WSN lifetimes. The root cause of this phenomenon is networking of nodes. In practical deployments, there are only few locations, where sinks can be positioned. In this case, there is always a set of nodes that are more involved in packets routing and forwarding. Since the lifetime of a network is defined as the time until there is no path to the sink, we see that uneven energy consumption places severe constraints on lifetime. One way to avoid unequal energy consumption is to get rid of networking. Mobile sinks may allow achieving this goal. To avoid human involvement the collection of data must be completely automatic. The obvious choice would be to use unmanned aerial vehicles (UAV)[40, 41]. The use of UAV as a mobile WSN node for emergency applications has been suggested in [42]. The authors in [43] proposed to use UAV for charging and deploying WSN nodes. Nevertheless, to date, no detailed investigations of such solutions and/or their comparison with conventional routed WSN designs have been performed. In our study, we propose a new UAV-assisted solution for data collection in WSNs. We optimized performance of single-hop communications between a sensor node and UAV in terms of optimal altitude and flying speed. We also compared lifetimes, coverage and required density of nodes of our solution with those of routed WSNs designs. We considered a wireless sensor network composed of multiple communicating entities equipped with the Bluetooth Low Energy (BLE) [44] transmission devices and UAV available to move in a certain area (see Figure 3.1.1 for the proposed structure).

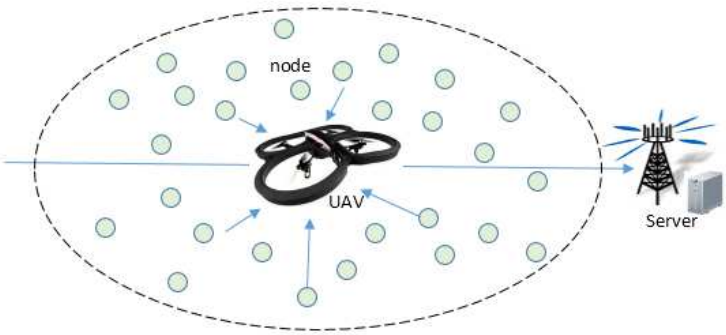


Figure 3.1.1: Proposed WSN data collection mechanism.

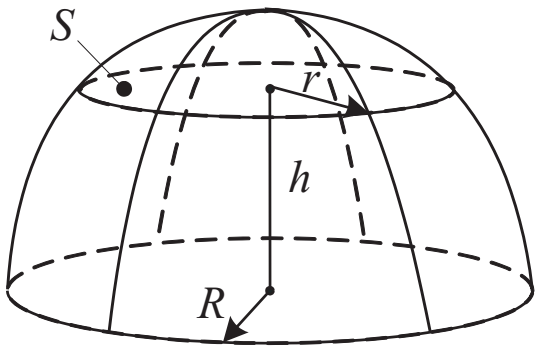


Figure 3.1.2: An illustration of the UAV flying over the BLE sensor node.

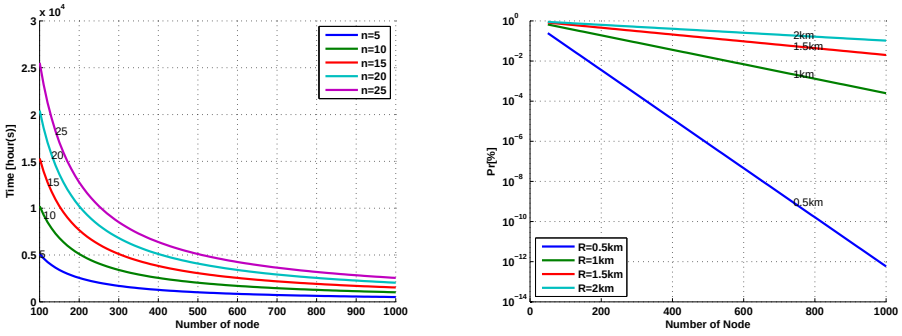
The proposed WSN system is based on single-hop data transmission approach where a UAV carries a sink node, BLE master, and fly over the monitored area to collect data from deployed sensor nodes, BLE slaves. Nodes communicate with each other using BLE. Sink node continuously scan for BLE slave nodes to establish connections and exchange data. After data collection, UAV comes back to ground station. The system does not require time synchronization between UAV flying time and BLE slave wake up pattern.

To exchange data BLE device have to discover its neighbors. It starts with entering the searching state and then proceeding with connection state. In searching state, the sensor node sends packets over three designated channels (37, 38, and 39). The scanner scans these channels continuously. The role of the scanner is performed by the UAV. When scanner discovers that searching node, it sends a connection request packet to establish connection. Once replied, both devices enter the connection state. While connected, searcher becomes slave and scanner is designated as master. In the system, transmission time is limited between *2 ms* and *14 s* in order to guarantee transmission of the required data. We also consider that the UAV is flying over the sensor node assuming that the antenna is omnidirectional forming a half-sphere around the sensor node's position as shown in Figure 3.1.2.The typical speed of UAV varies in the range of 30-45km/h, 10-15m/s. BLE range, R, is assumed to be upper bounded by 100m. We assume that when UAV crosses the half-sphere it goes through the center of the cutting plane to increase its chances to establish connection and collect the data from the sensor.

3.1.2 Performance analysis

In our system, we assume that all the sensor node have similar power supply. In our research, we calculated required overall consumption per transaction and got that our network lifetime is about 2020 days [P2]. While comparing proposed system and traditional (routed) WSN we consider as an example 100 sensor nodes, 25 sinks and data collection interval 10s, and as we can see in Figure 3.1.3(a) the maximum lifetime for that network is approximately 2552 hours (106 days) which is 18 times smaller than for the proposed system.

The fraction of uncovered area as a function of the number of uniformly



(a) Routed WSN lifetime (b) Fraction of uncovered area

Figure 3.1.3: Performance comparison of the proposed and conventional designs.

deployed nodes[45, 46] is shown in Figure 3.1.3(b). The sensing radius of nodes are assumed to be 100m while the monitoring area radii are 0,5; 1; 1,5; 2km. The fraction of uncovered area decreases exponentially when the number of nodes increases. Using these data one can estimate the number of nodes required to cover a certain area such that only a given small fraction on area is unmonitored.

In work [P2] we also analyzed the performance of the proposed design in terms of the coverage properties and compared to the conventional routed WSN design revealing that the density of nodes required to ensure coverage is approximately two times smaller compared to routed WSNs even when sensing region of nodes coincides with their communications range. These properties make the proposal an attractive option for monitoring environmental parameters in large open areas and might be used as one of Big Data systems collection techniques.

3.2 Bacterial networks

Intra-body sensor bio-inspired communications networks may open up new horizons for different monitoring systems (e.g. health monitoring, in-body food delivery etc.). A number of the studies propose deoxyribonucleic acid (DNA) to pre-encode the data in such systems. As the end systems may not be powerful enough in terms of the processing power to extract the information from the encoded DNA the amount of data generated by such systems could be large making them a potential sources of Big Data.

The use of flagellated bacteria to carry DNA-encoded data is one of the enabler techniques for prospective nanonetworks. There are three major properties making the flagellated bacteria useful for communications. First, they have the ability to store DNA-encoded data in chromosomes or plasmids [47]. Bacteria are also able to pick up swimming DNA particles through the process of transformation [48]. Finally, the inherent ability to swim in the environment may potentially enable delivery of encoded messages to a communicating party, e.g. a receiver nanomachine [49].

The nature of the bacterial networks is completely different from the traditional electromagnetic communication imposing several restrictions and challenges which will be discussed and analyzed in this chapter together with performance analysis of bacterial communications. Particularly, here, we develop

an analytical framework for single and multiple transmitter (Tx) and receiver (Rx) pairs communicating in a closed environment. More detailed description of the analytical framework can be found in [P3].

3.2.1 System model

For communications in bio-inspired nanonetworks, the scientists often use Escherichia coli (E.coli) bacteria as it is widely studied and is the least harmful. Ability of bacteria swimming in the environment potentially allows it to deliver some data [49]. The concept of bacterial networks was proposed and studied in single-hop and multi-hop settings in a number of studies, addressing metrics such as capacity, communication range and delay [50, 51, 52]. The common accepted communications entities and the associated information delivery process is illustrated in Figure 3.2.1.

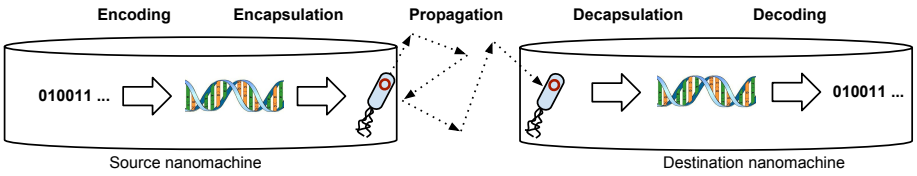


Figure 3.2.1: End-to-end bacteria communication model.

We can distinguish the following phases of the information delivery process:

- encoding of data;
- encapsulation;
- channel propagation;
- decapsulation;
- decoding.

First, message is encoded into the DNA strand at the transmitter [47]. At the second stage the information in terms of DNA strands is picked up by bacteria via the so-called transformation process [48]. Depending on whether a single or multiple bacteria will be emitted by a nanomachine it could also be replicated to enable multiple bacteria carrying the same information. Replication could be achieved via supplying food and waiting for a certain amount of time for bacteria to replicate. Alternatively, a bacterium could be injected in a compartment having other bacteria inside. In this case the information is replicated to a number of bacteria using the conjugation process which description can be found in work [P3]. At the receiver these operations are performed in reverse order, that is, we first decapsulate the information and then decode it. Our interest is in channel propagation stage which is exactly delivery of information via bacteria movements in a certain compartment.

As the information is delivered using randomly swimming bacteria the delay associated with data delivery could be substantial. Three following mechanisms have been proposed to speed up the process of data delivery:

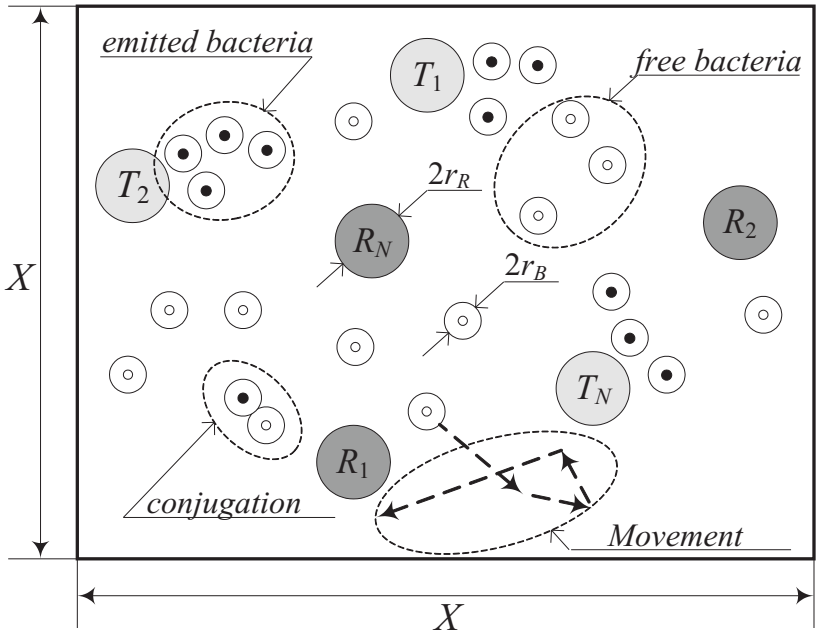


Figure 3.2.2: An illustration of the system model.

- chemoattraction;
- multiplication mechanisms;
- conjugation.

Detailed description of those mechanisms can be found in work [P3]. In our study, we have taken into account the latter two mechanisms.

Table 3.2.1 provides notation used in the model. We consider a squared $X \times X$ mm environment with N Tx-Rx pairs as shown in Figure 3.2.2. Positions of transmitters and receivers are assumed to be stationary and uniformly distributed over X^2. The length of all messages is identical and equal to L_M informational units measured in DNA base pairs. These messages are inserted into plasmids that are further injected into bacteria at the transmitter. To transmit a message each transmitter releases B_E, $B_E \geq 1$, bacteria containing the identical copy of a message into the environment. We also consider the case of N_R, $N_R \geq 1$, receivers for a single transmitter. We also assume that the size of transmitter and receiver are significantly larger than the size of E.Coli (2μm in length by 0.5μm in width) such that the latter can be effectively represented by a point. There are B_F bacteria moving in the environment. We call them "free" as initially they do not contain any information. They help to deliver the information to the receiver through the process of conjugation. The sensitivity radius of a bacterium r_B, $r_B << X$, is an area of circular shape. Once two bacteria are within the reach of each other the conjugation starts with probability p. During the conjugation the message is copied to another bacterium at rate $C = 800$ base pairs per second [48].

Table 3.2.1: Notation of the bacteria network model.

Parameter	Meaning
X	Side of a compartment
N	Number of Tx-Rx pairs
B_E	Number of emitted bacteria
B_F	Number of bacteria in environment
N_R	Number of receivers per transmitter
r_R	The radius of receiver
r_B	Sensitivity radius of a bacteria
p_C	Probability of conjugation
C	Conjugation rate (pairs per second)
L_M	The length of a message
μ	Rate of vibration process
υ	Bacterium swimming speed
τ	Bacterium straight swimming time
α	Bacterium tumbling angle

The system is subject to environmental "vibrations" caused by, e.g., ultrasound. Since the actual characteristics of the vibration process depend on many external and internal parameters, see e.g. [53, 54], we model it using the homogeneous Poisson process with intensity μ. Once the system is shaken the conjugation stops abruptly and bacteria continue to move randomly. The message is considered to be delivered when a bacterium carrying a full copy of the message reaches the receiver. The sensitivity radius of all receivers is assumed to be r_R, $r_R >> r_B$. We are interested in delay-related metrics of the message delivery process including cumulative distribution function (CDF), mean and 0.95-quantile.

3.2.2 Performance analysis

In classic electromagnetic communications systems one of the most important performance metrics is the delay between message transfers between two communicating entities. For the bacteria systems with free bacteria in the environment used for data transmission propagation delay is a random variable. For single bacteria, it is described by the so-called first-passage time (FPT) to a receiver. The FPT gives the time of a first contact between a bacterium and the receiver. As a result, it provides the delay performance of a communications link. Since conjugations are allowed in our model we also need FPT distribution for free bacteria. Finally, due to possibility of multiple conjugations another time-related metric we need is inter-passage times (IPT) between two bacteria. It is defined as the time between two successive meetings of bacteria.

As we described in the model there are N Tx-Rx pairs and in order to analyze performance we considered two possible cases: we proposed model for single Tx-Rx pair and we developed model for multiple Tx-Rx pairs with and without usage of free bacteria for data transmission. The major part of the research is consist of not just model development but of numerical results we obtained and comparison of them which is described in detail in [P3].

There is a description of the model for single Tx-Rx pair. Assume that at

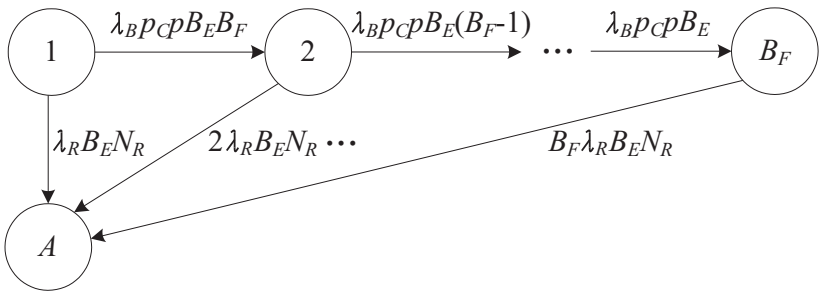


Figure 3.2.3: Absorbing Markov chain model for single Tx-Rx pair.

time $t = 0$ the transmitter emits B_E bacteria each having a message of length L_M. λ_B is a rate of FPT exponential distribution between bacteria,λ_R is a rate of FPT exponential distribution between bacteria and the receiver. As it described in [P3], one of these bacteria hits the receiver with rate $B_E\lambda_R$ ($B_E N_R \lambda_R$ for N_R receivers per transmitter) or may get engaged in the conjugation process with free bacterium. If the former happens then the process ends, that is, the message is delivered to the receiver. Since there is non-zero probability, $1 - p$, that two bacteria that come in contact with each other do not start the conjugation the process of conjugations is Poisson with rate $pB_E B_F \lambda_B$. The conjugation process may abruptly end due to external vibrations the system is subject to. These vibrations happen according to the Poisson process with intensity μ. Once the system is shaken all ongoing conjugations are aborted. We are interested in those conjugations that are successfully finished, i.e. the whole message has been copied.

The process of message delivery to the receiver can be modeled using the continuous-time absorbing Markov chain model, $\{S(t), t \geq 0\}$ with the state-space $S(t) \in \{1, 2, \ldots, B_F, A\}$, where the states $\{1, 2, \ldots, B_F\}$ are transient modeling the number of bacteria having the message by time t, $t \geq 0$, state A is absorbing one. The state transition diagram of the model is shown in Figure 3.2.3. From any state i there are only two possible transitions, to the state $i+1$ and to the absorbing state A. The rates out of a state i are

$$\lambda_{i,i+1} = ipp_C \lambda_B B_E (B_F - 1),\ \lambda_{i,A} = i\lambda_R B_E N_R, \tag{3.2.1}$$

where the p is probability of initiating conjugation and C is high (800 base pair per second [48]) .

For the multiple Tx-Rx case assuming that all transmitters start communicating at a certain time $t = 0$ releasing B_E bacteria each. We concentrate on a randomly chosen Tx-Rx pair and consider the rest of transmissions as a single "interference" process emitting $(N-1)B_E$ bacteria at $t = 0$ and competing for shared resources (free bacteria) with the transmission of interest. This pair is called tagged in what follows. The resulting model is a direct extension of the model for a single Tx-Rx pair. The state transition diagram is shown in Figure 3.2.4, describing the case of $B_E$emitted bacteria by each of N sources. The transitions in each row correspond to increasing the population of bacteria of the tagged source due to conjugation process. The Transitions in each column correspond to the increase of the population of the interfering process. Finally,

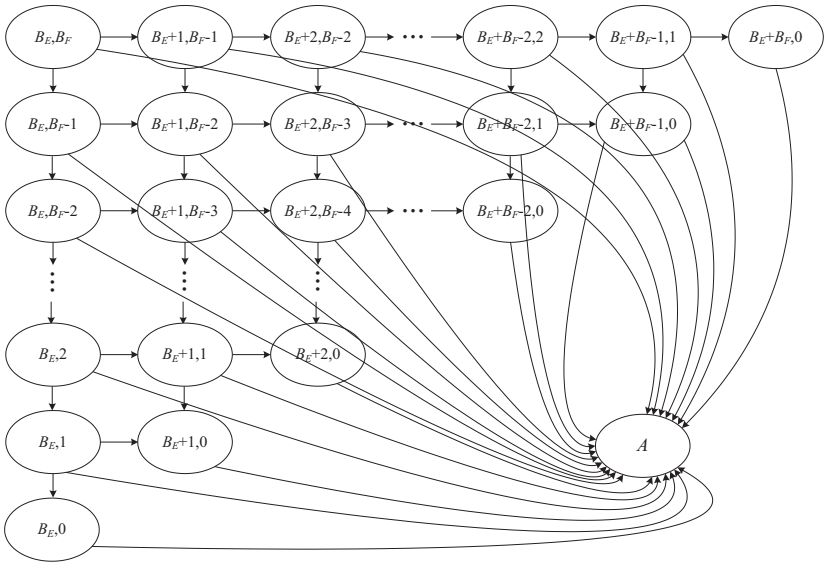


Figure 3.2.4: Absorbing Markov chain model for multiple Tx-Rx pair.

Table 3.2.2: Input parameters.

Parameter	Meaning
C	Conjugation rate, 800 base pairs/s, [16]
ν	Swimming speed, 20μm/s, [3]
τ	Mean straight swimming time, 3.5s, [3]
α	Tumbling angle, *U(0, 2π)*, [3]
$1/\mu$	Mean of vibration process, $1/\mu = 0.01$s
r_R	The radius of receiver, 50μm
r_B	Sensitivity radius of a bacteria, 5μm
p_C	Probability of initiating conjugation, 0.7
L_M	10 base pairs

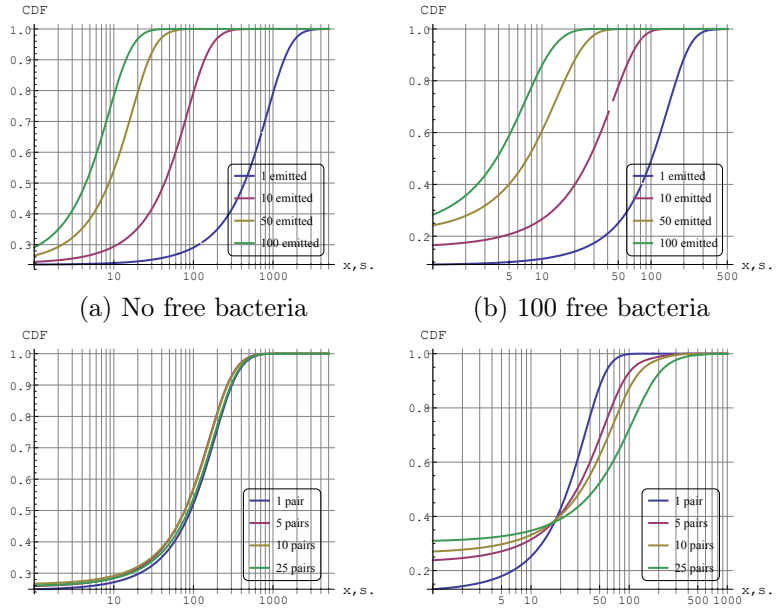
transitions to the absorbing state are possible from any state of the chain.

3.2.3 Numerical insights

In order to get numerical results obtained using the models formulated in previous section, we defined input parameters which are listed in Table 3.2.2 and the performance metrics of interest are CDFs, mean and 0,95- quantile of message delivery time.

Detailed description of the experiments and numerical results are presented in [P3]. Some obtained numerical results of CDFs of delivery time for different input parameters, mean delivery time and 0,95- quantile of message delivery time for single and multiple Tx-Rx pairs are shown in Figures 3.2.5,3.2.6, 3.2.7 (in all Figures (a) and (b) for single pair, (c) and (d) for different numbers of pairs).

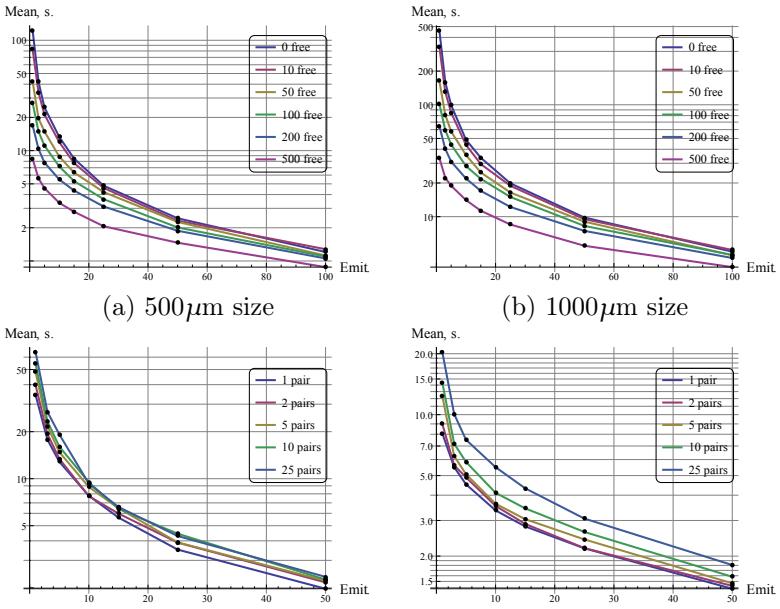
We performed a systematic investigation of the effect of input parameters



(a) No free bacteria (b) 100 free bacteria

(c) No free bacteria, 1 emitted (d) 100 free bacteria, 1 emitted

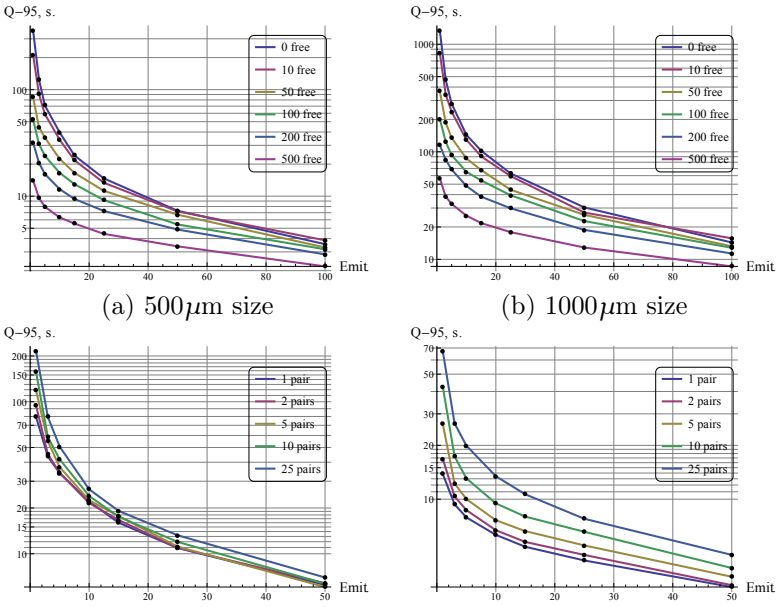
Figure 3.2.5: CDFs of delivery time for different input parameters.



(a) 500μm size (b) 1000μm size

(c) 500μm size, 50 free bacteria (d) 500μm size, 500 free bacteria

Figure 3.2.6: Mean delivery time for different compartment sizes.



(a) 500μm size (b) 1000μm size

(c) 500μm size, 50 free bacteria (d) 500μm size, 500 free bacteria

Figure 3.2.7: 0.95-quantile of delivery time.

on the delivery performance of the system. The following are the main findings: (i) both the number of emitted bacteria and the number of free ones provides noticeable quantitative gains in terms of mean delivery time, (ii) the gain provided by the number of emitted bacteria is significantly stronger compared to the effect of free bacteria, (iii) the relative gain of free bacteria strongly depends on the compartment size (iv) the relative gain associated with the number of emitted bacteria is independent of the compartment size. The effect of multiple Tx-Rx pairs is straightforward as the increase in their number leads to the corresponding increase in the mean delivery time. Importantly, we provided the results for the 0.95-quantiles of delivery time that can be used in practice to decide upon the number of emitted and free bacteria such that a delivery time bound is satisfied. Novel bacteria-based networks may not be powerful enough in terms of the processing power to extract the information from the encoded DNA, but the amount of data generated by such systems could be large making them a potential sources of Big Data.

3.3 Data from users for data collection networks

As an example of the novel WSN at the macro-scale we have proposed and analyzed in detail a flash-crowds monitoring system allowing to gather aural and visual information in real-time. The feasibility of such systems is based on the fast growing usage of cellphones and mobile telecommunication networks. We have mainly concentrated on efficient data collection from users, particularly, their devices equipped with special sensors as a part of unified monitoring system. For the proposed system we have analyzed the coverage metrics and estimated the rate imposed on the wireless network showing that modern sys-

tems may not support a detailed visual monitoring. More detailed description of the modeling and evaluation is presented in [P4].

3.3.1 System model

For the crowd monitoring system we should define the environment and characteristics of the system. First of all, while talking about such systems we assume them as systems without fixed infrastructure. We further assume that those areas can not be covered with the UAVs as we proposed in Section 3.1. As we were focused on users with their sensing devices (e.g. cellphones) equipped with microphones and cameras we proposed audio and video monitoring assuming that user could engage the monitoring process by downloading special application on the mobile device. As we cannot guarantee all the users join the monitoring system we assumed the uniform distribution of users over the monitored area. We obtained coverage metrics including the CDF of the covered area, mean and quantiles for both aural and visual information. Another important issue of the system model that we considered was blocking of camera by other users (humans) located in the area [55]. For the proposed system, local information processing may not be useful as a single node may not have enough of data to make conclusive decision. Indeed, the strength of the proposed system is in the ability to get information from many sources located nearby. Thus, the information shall be delivered to the certain remote server for further centralized data processing. The devices participating in the monitoring process are expected to use the resources of cellular system uploading the data to the remote server. The problem was formalized as follows: for a random placement of users on the landscape what should be the density of nodes providing coverage for a certain type of media such that a percentage of area is covered with probability of x. We also were interested in the amount of wireless network resources needed to monitor the area of certain dimensions.

Since the height of user devices is assumed to be comparable with the height of blockers (humans) we limited our interest to two-dimensional scenario. Fixing a certain time instant t we got and illustration of the system in Figure 3.3.1.

The area to be covered was assumed to be 100 by 100 meters. The humans were represented by circles on the landscape of diameter d. There are overall $N+M$ humans in the area comprising a crowd to be monitored. N humans were assumed to follow a conditional Mattern process with parameter d in the area [56, 57]. In other words, no two users could be closer than at the distance $2d$ to each other as in practice human bodies do not overlap. M additional humans participate in the monitoring process and they also follow conditional Mattern process with parameter d. Thus, the overall number of potential blockers for viewing field of cameras is $M + N$. In the model we also considered different types of sensors - audio and video with different coverage depending on the sensor type. For audio sensors, such as microphones, the assumption of circular coverage with radius r_A around a users is taken as humans do not block acoustic waves propagation significantly [55]. For visual sensors, such as cameras capturing video or still images, the field view is by default of sectoral shape with radius r_V . We model cameras as triangle with the height to the base r_V and apex angle α . To include a random orientation of cameras we assume that the bisect of the apex angle is uniformly distributed in $(0, 2\pi)$. Humans that fall into coverage field including those participating in the monitoring process block view as we

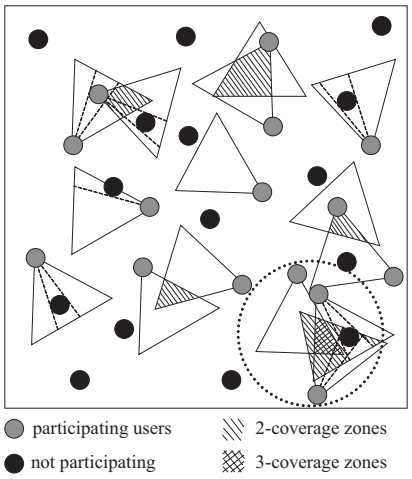
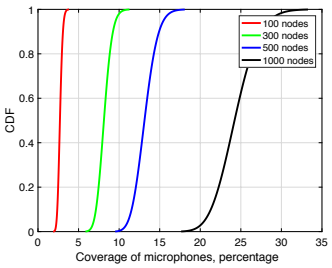


Figure 3.3.1: The illustration of visibility in the dense crowd.

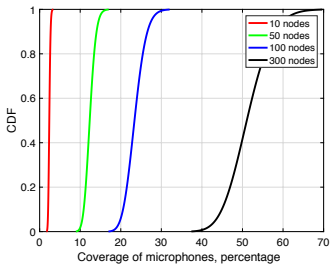
can see from the Figure 3.3.1.

3.3.2 Performance analysis

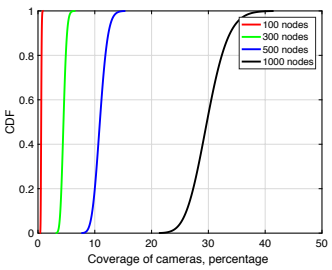
The coverage metrics have been obtained using simulation approach. The coverage CDFs for different number of participating users and different coverage radius of a single user are shown in Figure 3.3.2.



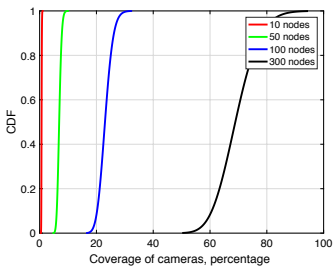
(a) Microphones, $r = 1m$



(b) Microphones, $r = 3m$

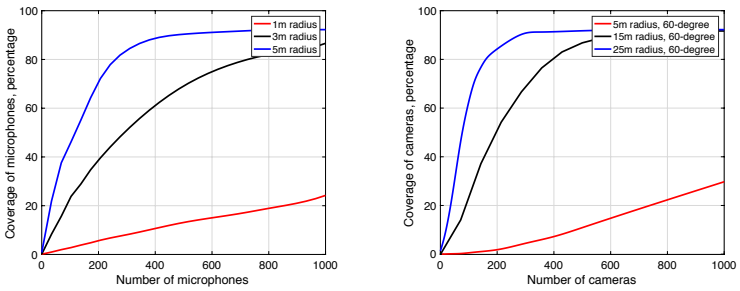


(c) Cameras, $r = 5m$



(d) Cameras, $r = 15m$

Figure 3.3.2: Cumulative distribution functions of coverage for microphones and cameras.



(a) Microphones (b) Cameras

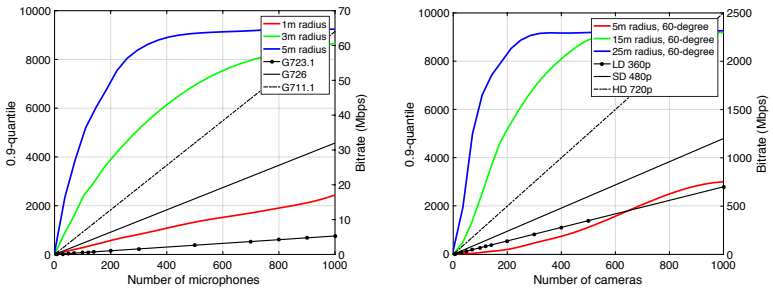
Figure 3.3.3: Mean coverage by microphones and cameras.

The number of non-participating humans was kept constant and equal to 1000. Note that instead of the absolute values we plot the percentage of the covered area in OX axis. Expectedly, for the same number of participating users better coverage is provided for larger coverage radius of a single node. Furthermore, increasing the number of participating users provides better coverage. However, as one may observe, even for extremely large number of participating users (e.g., 1000 nodes) full coverage is provided with negligible probability for aural information. Blocking of visibility field in visual information scenario does not qualitatively affect the form of CDFs compared to non-blocked aural information scenarios.

The mean values of the area coverage percentage as a function of the number of participating users and different coverage radii of a single user are shown in Figure 3.3.3. The number of non-participating users is set to 1000. One important behavior of this metric is that it does not approach 100% even for extremely high number of users and rather large coverage of single users (e.g., 25 meters for visual information). This behavior is attributed to completely random choice of the participating users (uniform distribution over the area).

Rate requirements are depend on the quality of the codec which is used for audio and video (detailed description of the codecs [58] can be found in work [P4]). The network requirements in terms of the bitrate needed from the network as well as 0,9 quantile of the coverage process are plotted in Figure 3.3.4 as a function of the number of participating users for different types of codecs.

Our numerical results indicate that the required density of the participating users needed to be exceptionally high to achieve “almost full” coverage (e.g., 0,9 quantile) for both audio and video sensors. Although the associated network requirements are exceptionally high they can be supported by the forthcoming millimeter wave or terahertz systems offering substantial rate boost at the air interface. Thus, crowd monitoring systems can be effectively used as a data collection networks but data delivery networks should be wisely chosen according to the requirements of the bitrate.



(a) Microphones, 0.9-Q (b) Cameras, 0.9-Q

Figure 3.3.4: Coverage quantiles and network requirements.

3.4 Big Data collection networks evaluation summary

In this chapter we have considered three different examples of the networks that could act as sources of Big Data. We have argued that the use of conventional WSNs is nowadays rather limited. Extending the lifetime of such systems using non-routed automated data collection techniques may significantly expand their usage. We then proceeded with bacterial systems analyzing the performance of the data collection techniques. Finally, we have proposed a new flash-crowd monitoring system that could potentially be an important source of Big Data. In the next chapter we analyze the process of data delivery from the WSNs egress points (e.g., sinks) to the Internet backbone (e.g., cloud data storage).

Chapter 4

Data delivery networks

In the previous chapter we argued that WSNs might potentially serve as individual sources of Big Data. To offload the traffic from such systems wireless access technologies are expected to be used. We have also demonstrated that modern wireless technologies might not be enough for this purpose.

In this chapter we concentrate on the analysis of THz and mm-wave wireless backbones serving aggregated traffic from WSNs. First, taking into account the specifics of these frequency bands including the highly directional antennas, molecular absorption and human blockage, we use the tools of stochastic geometry to obtain the estimates of aggregated interference. We further proceed formulating the model of the service process of packets on the wireless channel at the network layer using the queuing theory approach.

4.1 THz and mm-wave for data delivery networks

The ubiquitous use of various personal devices has dramatically increased the connectivity all around the world. Media sharing such as videos and audio, sensing data etc. have led to an extreme rise in the data transfer between devices. It is predicted [59]that in 2016, there would be more than 1 zettabyte of data being shared across the globe and by 2020 there will be 2.3 zettabytes. The future generation of wireless systems is expected to rely on small cells to offload the traffic from current cellular and local area networks. Going up in the frequency band from microwaves to mm-waves (30-300GHz) and THz (0.3-3THz) is one of the ways to principally increase the capacity of such small cells. In addition, to service human generated traffic, they could also be used to offload the data generated by the WSNs.

4.1.1 Basic characteristics

The THz and mm-waves region is defined from 30 GHz to 3 THz (Figure 4.1.1) based on the wavelength of the electromagnetic wave. Electronic devices have been mainly operating in the low frequency regime of this spectrum and their performance degrades as one approaches their cut-off frequencies. A significant proportion of the electromagnetic spectrum is unexplored starting

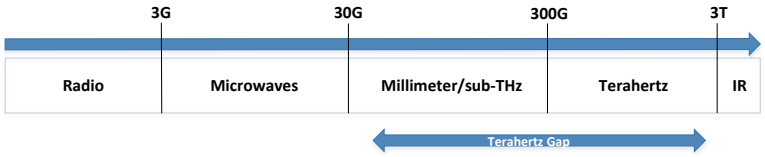


Figure 4.1.1: Electromagnetic spectrum showing the millimeter/terahertz region.

from 50 GHz to 3 THz and is popularly referred to as the terahertz gap. The reason for this name is that this band is too high for making efficient equipment to generate electromagnetic waves. The photonics techniques used for even higher frequencies are also inefficient for THz band,

Recently, there has been significant interest in the 60 GHz band for high data rate communication in both outdoor and indoor networks. The THz and mm-waves band is also becoming popular for imaging applications at 94 GHz, 140 GHz and 220 GHz [60]. Applications in the automotive radar industry in the 77-78 GHz band are gaining interest for blind spot detection to minimize accidents. Terahertz chemical imaging or molecular spectroscopy is another emerging area of application where certain substances can be detected based on their high degree of absorption at these frequencies. This can be used to detect harmful gases such as carbon monoxide (which has response at 230 GHz) or phosphine (which has response at 266 GHz). Some researchers are already investigating the use of 300GHz, 640GHz, and even the entire THz band [61, 62, 63].

Indoor wireless communications with THz may provide multiple data channels with gigabit per second or greater capacity. The data bandwidth would exceed wireless protocols such as IEEE 802.11b, and the propagation distance, though limited, would be competitive with line-of-sight IR. Existing channel models for the Megahertz and the Gigahertz frequency bands cannot be reused for the Terahertz Band, because they do not capture several effects such as the attenuation and noise introduced by molecular absorption, the scattering from particles which are comparable in size to the very small wavelength of Terahertz waves, or the scintillation of Terahertz radiation [64]. Effectively any object whose size is greater than few millimeters acts as a blocker for such systems. At the same time, the size of the antenna will be reduced, but antenna should be high directive allowing to partially overcome severe propagation losses. Finally, there is also the low efficiency and relatively low power available from currently available sources.

Such technologies are the approach to improve communication between the end user and base stations and thus they are usually presented as a part of heterogeneous network. Usually that heterogeneous network consists of the mm-wave backhaul/fronthaul integrating small cells in the traditional cellular network. The small cell can have an access link compromising both conventional cellular access such as LTE and a novel millimeter-wave link utilizing a centralized radio access network [65]. While talking about Big Data systems applications taking into consideration different data collection networks we assume that mm-waves backhaul will be utilized as a delivery technology in our case delivering data from the data collection networks to the Internet backbone

(e.g. cloud data storage).

4.1.2 Challenges

For our purposes, there are network challenges, which should be further studied such as radio propagation and interference structure. Such performance analysis would help us to understand whether those networks are potentially our data delivery networks for Big Data systems or not. The requirements for an outdoor mm-waves channel model are expected to be very similar to the indoor case, which is well-described in IEEE 802.11ad documents [66]. The channel model should provide accurate space-time characteristics of the propagation channel, support beamforming with steerable directional antennas on both Tx and Rx sides with no limitation on the antenna technology, taking into account polarization characteristics of antennas and signals and support non-stationary characteristics of the propagation channel. High gain beamforming antennas are therefore needed at the small cell base station as well as the user terminal. The most straightforward solution that qualifies for all requirements is the mm-waves phased array antenna, which is successfully used for prototypes [67]. However, creation of such large-aperture antenna arrays may pose a problem due to production cost, heat dissipation and feed circuitry complexity. As it was indicated before for such systems blocking should be studied better. Detailed study on the mentioned challenges is presented in [P5,P6].

4.2 Wireless connection performance

While studying mm-wave and THz systems we consider similar radio propagation characteristics and critical factors for them. First, the most important critical factor there is that Line-of-Sight (LoS) signal in these bands can be effectively blocked by almost any obstacle. While reflections of the objects in the channel do contribute, their effect is of secondary importance and presence of LoS often dictates the channel quality [68]. The second inherent feature of the considered bands is molecular absorption. The phenomenon is related to absorption of electromagnetic energy by the molecules in the environment having resonant frequencies in the communications band of interest. In EHF band, oxygen, that is abundant in the atmosphere, affects the path loss. The net result is additional 10-20dB loss per kilometer [68]. There are different types of such absorbent. As a result power at the receiver is described by more complex expression. The last critical point there is a high antenna directivity which from one point allows to compensate for propagation losses and from another point it produces very narrow and precise beam. To understand the trade-off between the directivity of antennas, the requirements imposed on beamsteering system and the signal-to-interference ratio (SIR) or signal-to-interference-plus-noise (SINR) interference models for directional antennas are needed.

In our research we considered several antenna models: cone model and cone-plus-sphere model. In the first model, the directivity of the transmitter is taken into account considering the coverage zone to be of cone-shape. This model is an abstraction assuming no side lobes and constant power at a certain separation distance from the transmitter. The second model takes into account imperfect antenna radiation pattern by modeling side lobes as a sphere around

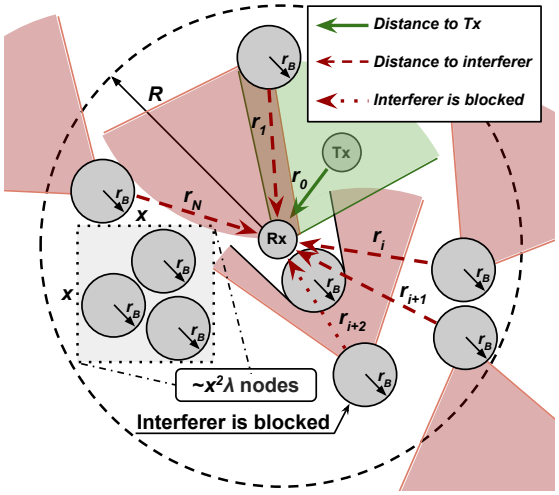


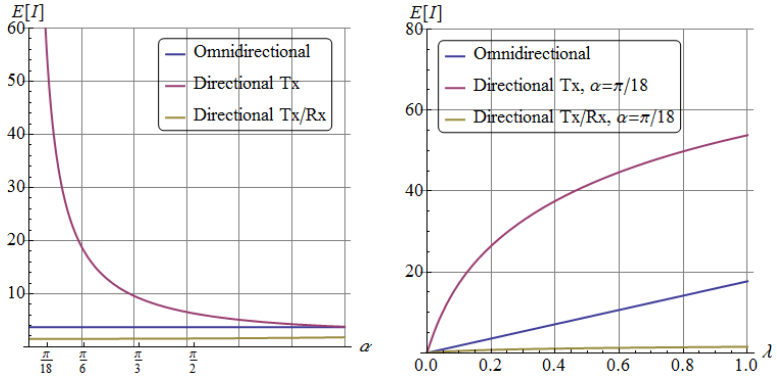
Figure 4.2.1: An illustration of the considered network deployment.

the antenna. Still, the power of the main lobe is assumed to be constant and depend on the distance from the antenna only. To parameterize the cone antenna model, we need to provide coefficient A corresponding to a directivityα of the antenna. For the second model, A_1 and A_2, corresponding to the main and side lobes, respectively, have to be provided. Coefficients A, A_1, and A_2 are used in the propagation model to properly receive the signal with respect to the direction it goes to or comes from.

While analyzing the network we considered random nodes deployment as it is shown in Figure 4.2.1 and only ideally directed towards each other antennas.

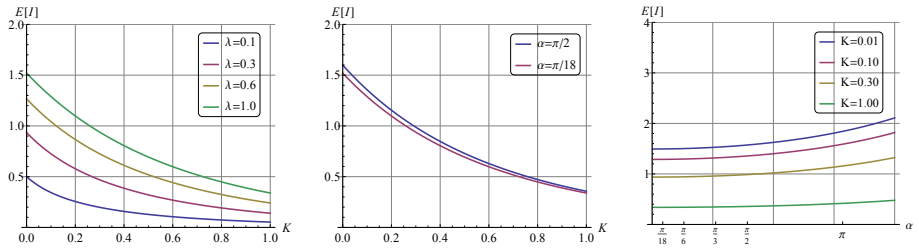
For the described scenario we were interested in the mean value of interference observed at the receiver. We also considered blocking scenarios which are in details described in [P5, P6]. α- is antenna directivity angle, and λ- is the intensity of interference/blocking in the area. As a result in Figure 4.2.2(a) we observe that the system with directional Tx or Rx antennas performs better starting from antenna directivity $\alpha = \pi/6$. Similarly, the interference ($E[I]$) for system directional Tx and Rx reduces even further for all considered values of α. The effect of interferers intensity on the mean interference ($E[I]$), shown in Figure 4.2.2(b), illustrates that, due to the effect of blocking, the large values of λ may lead to better performance when directional antennas at Tx or Rx only are used. The reason is that the system without omnidirectional antennas is characterized by the linearly growing interference in presence of blocking, while the system with directional antennas – by logarithmically growing interference.

Having identified significantly better performance of a system with directional Tx and Rx, from now on we concentrate on this system. Let us first illustrate the effect of absorption coefficient. Figure 4.2.3 highlights dependence of the mean interference ($E[I]$) on the absorption coefficient K for cone directional antenna model with blocking (where value of the K in between 0 and 1). Fixing the density of interferers, λ , we observe the expected dependency K, i.e., the interference is smaller for higher values of K, see Figure 4.2.3(a). In general, when K increases, the interference naturally decreases due to less radiation reaching the receiver. It is important to note that this feature of THz



(a) Blocking, varying α, λ= 1 (b) Blocking, varying λ

Figure 4.2.2: Comparison of interference for scenarios with omnidirectional and directional (cone) models.



(a) Varying K, α =π/12 (b) Varying K, $\lambda = 1$ (c) Varying α, $\lambda = 1$

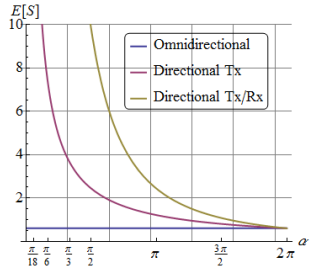
Figure 4.2.3: Dependence of the mean interference on the absorption coefficient K for cone directional antenna model.

band is often claimed to have negative effect rather than positive. We also see that proper choice of the emitted power and the operational frequency may, in fact, allow for point-to-point links creating only little interference to concurrent transmissions.

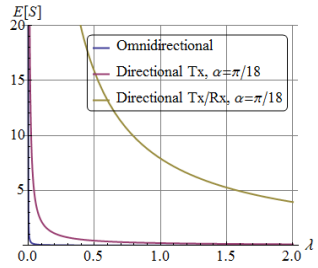
The interference alone does not allow to make ultimate conclusion about the performance of THz systems. The reason is that antenna directivity not only affects the interference but the useful received signal strength too. Thus, the study of interference in [P5] has been extended to characterize SINR in next paper. Below, we assess performance of the considered scenarios using SINR as a metric of interest concentrating on the cone antenna model (r_0 is the distance between Rx_0 and Tx_0).

As one may observe from Figure 4.2.4(a) the system with directional Tx and Rx demonstrates the best performance. However, either Tx or Rx side is equipped with directional antenna SINR increases as α decrease. Recall that the aggregate interference in this case also increases as α gets smaller. However, it is compensated by the increase in the useful received power.

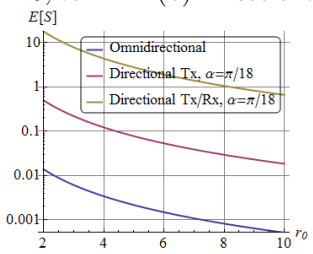
Figure 4.2.4(b) highlights that increasing the density of interferers results in corresponding exponential decrease of SINR. Again, the system with directional Tx and Rx greatly outperforms the system with directional Tx or Rx. The worst performance is observed for omnidirectional antennas. Finally, Figure 4.2.4(c)



a) Effect of α, r_0 = *3*, λ= *1*



(b) Effect of λ, α =π/18, r_0=*3*



(c) Effect of r_0, α =π/18, λ= *1*

Figure 4.2.4: The effect of antenna directivity on SINR.

shows the effect of the distance between Tx and Rx, r_0. The SINR decreases in r_0 for all considered models. The system with directional antennas at both Tx and Rx outperforms the one with omnidirectional one by two orders of magnitude.

When directional antennas are used at both Tx and Rx, the interference drastically decreases and, overall, the SINR is much improved too. The inherent property of THz bands of self-blocking of radiation by interferers leads to drastic performance improvements in terms of aggregate interference and SINR metrics compared to microwave systems. Similarly, molecular absorption can also help to increase the SINR. While molecular absorption loss further decreases the received signal exponentially, it decreases interference too, and the effect on the latter is more profound. In our study we assumed that interferers may completely block the THz radiation. In practice, reflections as well as scattering of electromagnetic waves inherent for these frequencies may still contribute to the aggregate interference at the Rx even when LoS is blocked requiring more advanced analysis. Novelty of the presented approach consists of research and performance analysis of the blocking effect when surely influence of directional antenna has been known before.

4.3 Backhaul performance analysis

The potential capacity of wireless backhauls operating in either mm-wave of THz frequency bands may allow for multiplexing of a number of arriving flows from different WSNs for further transmission over the wireless channel to the Internet access point (e.g. cloud data storage). Since the individual traffic streams can be heterogeneous in nature and may have different service requirements, their individual performance is of special interest. Per-source analysis of heterogeneous sources is non-trivial as contribution of each multiplexed source

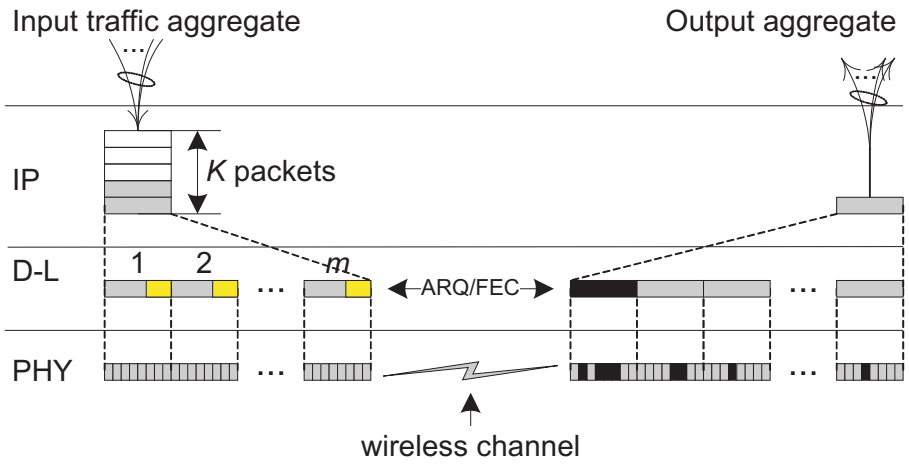


Figure 4.3.1: The system model of the backhaul link.

needs to be explicitly accounted for. It becomes even more complicated when individual traffic flows are auto correlated.

In our study we concentrated on performance parameters provided by a wireless channel to a single auto correlated traffic flow in presence of a certain number of concurrent auto correlated flows of the same priority. In our work [69] we have demonstrated that the packet service process of the wireless channel with forward error correction (FEC) and non-persistent (truncated) automatic repeat request (ARQ) can be sufficiently well represented using G/G/1/K framework with independent identically distributed (iid) service times. To capture auto-correlation in the input traffic flows we have used a general discrete-time batch Markovian arrival process (D-BMAP). The system of interest was then modeled as D-BMAP+D-BMAP/G/1/K queuing system with two concurrent processes of the same priority, where the first process is the arrival process from the source of interest (tagged source) while the second one represents superposition of one or more concurrent sources. For both superposed and tagged arrival processes we provided expressions for probability functions (PFs) of the number of lost packets in a slot and delay experienced by an arbitrary packet. Since losses in our model were allowed to occur due to both buffer overflow and imperfect error correction we further obtained expressions for PFs of the aggregated number of lost packets for both tagged and superposed arrival processes.

The high-level abstracted view of the service process on a wireless backhaul is represented in Figure 4.3.1. Note that the presence of the autocorrelation in the arrival flows significantly complicated the analysis of the queuing model leading to the complex framework of D-BMAP+D-BMAP/G/1/K queue. Thus, the major question of interest was the impact of the autocorrelation on loss and delay metrics experienced by the tagged flow.

We assume that a certain number of flows of the same priority, that may themselves be traffic aggregates, share a wireless link of the raw capacity. We assume that the size of all packets is constant and equal to N bytes including all headers. The buffering is done at the IP layer. The number of waiting positions in the buffer is limited to *K-1* packets. When there is at least one packet in the

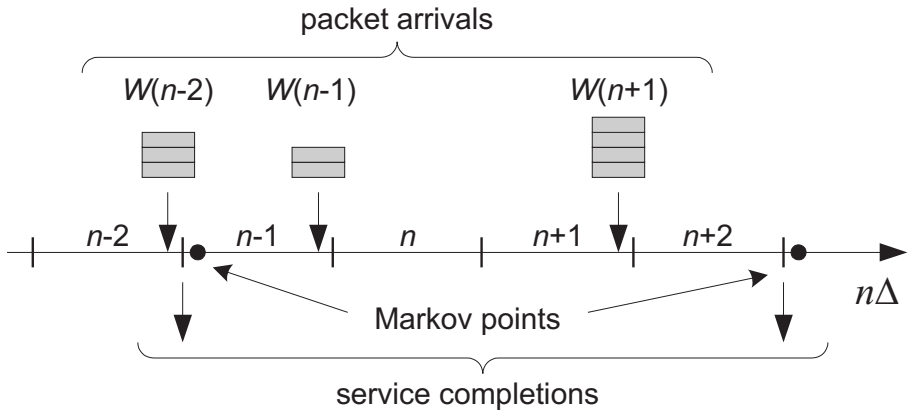


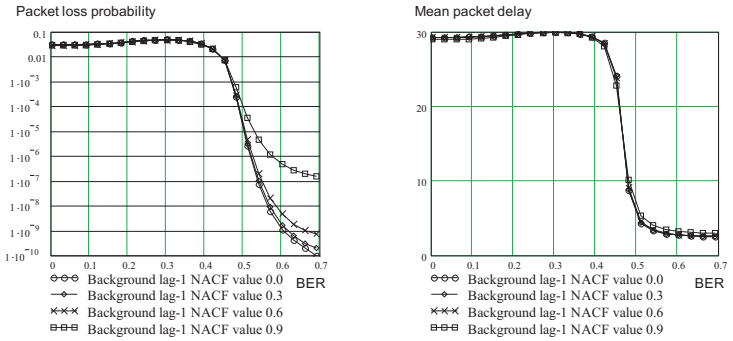
Figure 4.3.2: Time diagram of D-BMAP$_A$/G/1/K queuing system.

buffer and the channel is free for transmission this packet is scheduled to the data-link layer. Between these two layers packets are segmented into a certain number of frames. We assume that the protocol data units (PDU) of the ARQ protocol consist of exactly one code word and refer to them as frames. Then, a certain FEC code with the symbol length of m_S bits that can correct up to l incorrectly received symbols is applied and these frames are further scheduled to the ARQ engine. The frame size is assumed to be constant and equals to m_F symbols. We assume that the probability of undetected error is negligibly small as it decreases exponentially when the size of a frame gets larger [70].

To evaluate performance of a wireless channel with unreliable data-link layer we distinguished between those losses caused by the excessive number of retransmission attempts performed at the data-link layer and those resulting from the buffer overflow at the IP layer. We further proposed the queuing system describing the process of packet transmissions at the wireless channel and solve it for the superposed and per-source performance parameters. Losses caused by excessive number of retransmission attempts and those occurring as a result of overflow at the IP layer were combined together to obtain a single loss descriptor.

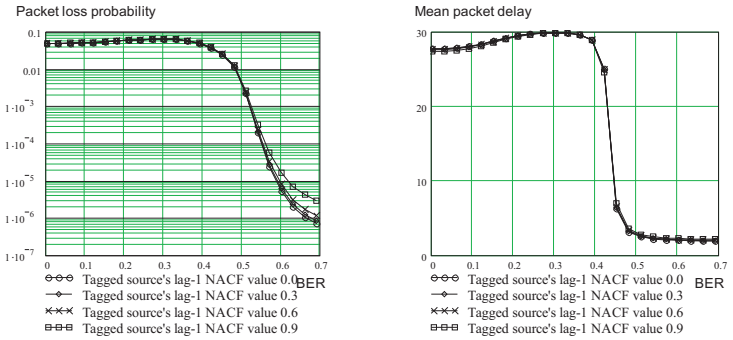
Time diagram of D-BMAP$_A$/G/1/K queuing system, where K is the capacity of the system measured in IP packets, *W(n)* - arrival processes) is shown in Figure 4.3.2. According to this system, packets arrive in batches, batches of packets arrive just before the end of slots. Arrivals are not allowed to seize the server immediately and the service of any arrival starts at the beginning of a slot. Packets depart from the system at the slot boundaries, just after batch arrivals. The state of the system is observed just after departure (if any). This system is known as the late arrival model with delayed access [71, 72]. The sojourn (service) time is counted as the number of slots spent by a packet in the system. More detailed description of the system is available in work [69].

Having the effect of autocorrelation in the tagged and background flows in mind, the selected results are presented in Figures 4.3.3 and 4.3.4. In the thesis the packet loss probability and delay in the system are presented as the function of BER which is non-trivial characteristics rarely seen in research papers. It is



(a) Packet loss probability (b) Mean packet delay

Figure 4.3.3: The effect of the lag-1 NACF of the background arrival process.



(a) Packet loss probability (b) Mean packet delay

Figure 4.3.4: The effect of the lag-1 NACF of the tagged arrival process.

required to further define influence of correlation on service characteristics.

Figure 4.3.3 shows the effect of the of lag-1 autocorrelation value of the background process for two values of the offered traffic load from the background process, ρ_B. Other parameters were as follows: the offered traffic load from the tagged process is $\rho_T = 0.6$, lag-1 autocorrelation of the tagged process is $K_T(1) = 0.0$, coefficient of variation of the number of arrivals from the tagged and background processes are $CV_T = 1.0$ and $CV_B = 1.0$, respectively. Note that the effect of the autocorrelation is limited, especially, for overflow-dominated regime of a channel, where most of the losses are caused by the buffer overflow not the wireless medium. Some differences are observed when the system is in impractical wireless-dominated regime, where the losses are mostly due to unsatisfactory wireless channel conditions. Our results (not shown here) also show that changes in the mean arrival rate of both processes do not qualitatively affect the effect of autocorrelation.

The effect of autocorrelation of the tagged process is shown in Figure 4.3.4. These illustrations have been obtained assuming $\rho_T = \rho_B = 0.6$, lag-1 autocorrelation of the tagged process set to $K_B(1) = 0.0$. In practical overflow-dominated regime autocorrelation does not affect performance metrics of interest significantly. In wireless-dominated regime some impact is observed. Comparing these

data to those shown in Figure 4.3.3 we observe that for the same input statistics the effect of autocorrelation in the tagged flow is less dominant. Changing rate of both processes does not qualitatively affect the response of the system.

Our detailed performance analysis is described in [69]. As a main result it is necessary to highlight that most losses are caused by the buffer overflow while after this point losses caused by imperfect error correction increase exponentially fast and dominate the loss process. In practical overflow-dominated regime performance of both aggregated arrival process and single arrival process in presence of concurrent traffic of the same priority is mainly affected by mean and variance of arrival processes. The effect of autocorrelation is almost unnoticeable except for extremely correlated packet arrivals. However, even in this extreme case the effect is at least one order of magnitude less than that of the mean value. For those studies when extreme accuracy is not required uncorrelated arrival processes can be used instead.

4.4 Big Data delivery evaluation summary

In this chapter we have analyzed the two aspects of data offloading from the network sources of Big Data using mm-wave and THz backhauls. We first demonstrated that the communications systems in these bands are characterized by the completely different interference structure and under certain conditions might operate in noise-limited regime as opposed to interference-limited regime of most microwave wireless communications systems. We then proceeded analyzing the service process of a single traffic flow in presence of concurrent flows of the same priority of a wireless backhaul. Our results showed that relatively simple queuing models can be used to obtain delay and loss metrics of the service process. Taken together, our results showed that that mm-wave and THz backhauls could serve as reliable wireless backhauls not only for 5G human-centric mobile systems but can serve as efficient technology for data offloading from networks sources of Big Data.

Chapter 5

Conclusions

Although the WSNs are widely considered as one of the sources of Big Data, the problems associated with these systems have not been deeply addressed in the past. One of the reason is that, it is the number of networks that are conventionally considered as contributors to Big Data not the individual networks. Motivated by the recent progress in conventional and principally new WSNs the presented thesis have concentrated on individual networks acting as sources of Big Data.

We have identified the following two critical issues related to such systems:

- how to efficiently collect the data in WSNs;
- how to efficiently deliver the collected data to the Internet backbone (e.g., cloud data storage).

While addressing these questions the following *main* conclusions have been made:

- We have argued that the micro/nano WSNs including both electromagnetic and bio-inspired ones could serve as individual sources of Big Data. The reasons are the specifics of encoding techniques (e.g., DNA-encoding) that induce a significant transmission overhead and the required simplicity of end devices that may not be capable of extracting information from the encoded data.
- The routed nature of conventional macro-scale WSNs significantly limits their application scope preventing from fast explosion of such systems in various industrial sectors. We have shown that the use of new data collection mechanisms may principally extend the lifetime of WSNs opening new markets and eventually leading to drastic increase of the generated traffic.
- We have demonstrated that the capacity of modern cellular systems can be insufficient to cover the needs of prospective WSNs applications. A set of new wireless technologies operating upper in the electromagnetic spectrum, such as millimeter wave or terahertz bands, is needed.
- The interference structure in millimeter and terahertz frequency bands is fundamentally different from that of microwaves with highly directional

antennas, molecular absorption, and human blocking affecting the resulting picture. We have shown that taking these factors together, wireless communications systems in these bands may operate in noise-limited regime with interference providing negligible impact on the overall system performance.

- Millimeter wave and terahertz systems may serve as wireless backhauls connecting the network sources of Big Data to the Internet backbone (e.g., cloud data storage). We have demonstrated that relatively simple queuing models can be used to assess traffic performance in such systems.

Bibliography

[1] S. Suthaharan, "Big data classification: problems and challenges in network intrusion prediction with machine learning," *ACM SIGMETRICS Performance Evaluation Review*, vol. 41, pp. 70–73, Mar. 2014.

[2] D. Evans, "The internet of things: How the next evolution of the internet is changing everything," *Cisco White Paper*, 2011.

[3] C. Snijders, U. Matzat, and U. Reips, "Big data: big gaps of knowledge in the field of internet," *International Journal of Internet Science*, vol. 7, pp. 1–5, 2012.

[4] G. Gallot, S. Jamison, R. McGowan, and D. Grischkowsky, "THz waveguides," *J. Opt. Soc. Am.*, vol. 17, pp. 851–863, Apr. 2000.

[5] D. Laney, "3d data management: Controlling data volume, velocity and variety." Gartner, 2001.

[6] B. B. Marr, *Big Data: Using SMART Big Data, Analytics and Metrics To Make Better Decisions and Improve Performance*. John Wiley and Sons, 2015.

[7] I. V., A. S., and R. S, "Security issues associated with big data in cloud computing.," *International Journal of Network Security & Its Applications (IJNSA)*, vol. 6, no. 3, pp. 45–56, 2014.

[8] J. Changqing *et al.*, "Big data processing in cloud computing environments.," in *Pervasive Systems, Algorithms and Networks (ISPAN)*, 2012.

[9] C. Jardak, P. Mahonen, and J. Riihijarvi, "Spatial big data and wireless networks:experiences, applications, and research challenges," *IEEE Network Magazine*, vol. 28, pp. 26–31, July/August 2014.

[10] I. Akyildiz, W. Su, Y. Sankarasubramaniam, and E. Cayirci, "A survey on sensor networks," *IEEE Communications Magazine*, pp. 102–114, 2002.

[11] F. Zhao and L. Guibas, *Wireless Sensor Networks: An Information Processing Approach*. Morgan Kaufmann Publishers Inc, 2004.

[12] C. Chong and S. Kumar, "Sensor networks: Evolution, opportunities, and challenges," in *IEEE*, vol. 91, pp. 1247–1256, Sept. 2003.

[13] S. Andreev, Y. Koucheryavy, N. Himayat, P. Gonchukov, and A. Turlikov, "Active-mode power optimization in ofdma-based wireless networks," in *Proc. GLOBECOM*, pp. 799–803, 2010.

[14] J. Horneber and A. Hergenroder, "A survey on testbeds and experimentation environments for wireless sensor networks," *IEEE Comm. Surveys and Tutor.*, vol. 16, no. 4, pp. 1820–1838, 2014.

[15] C. Chen, M. Won, R. Stoleru, and G. Xie, "Energy-efficient fault-tolerant data storage & processing in mobile cloud," *IEEE Transactions on Cloud Computing*, vol. 3, pp. 28–41, Mar. 2015.

[16] M. Komarov and D. Moltchanov, "System design and analysis of uav-assisted ble wireless sensor systems," in *Wired/Wireless Internet Communications* (L. Mamatas *et al.*, eds.), Lecture Notes in Computer Science, pp. 284–296, Springer International Publishing, 2016.

[17] I. Akyildiz, M. Pieborn, S. Balasubramaniam, and Y. Koucheryavy, "The internet of bio-nano things," *IEEE Communications Magazine*, vol. 53, pp. 32–40, Mar. 2015.

[18] F. Dressler, *Self-Organization in Sensor and Actor Networks.* John Wiley and Sons, 2007.

[19] M. Gaber, U. Roehm, and K. Herink, "An analytical study of central and in-network data processing for wireless sensor networks," *Information Processing Letters*, vol. 110, pp. 62–70, 2009.

[20] U. Roehm, M. Gaber, and Q. Tse, "Enabling resource-awareness for in-network data processing in wireless sensor networks," in *Proceedings of the nineteenth conference on Australasian database*, vol. 75, pp. 107–114, 2008.

[21] G. Peddibhotla, "Gartner 2015 hype cycle: Big data is out, machine learning is in." Online, 2015.

[22] Y. Hao *et al.*, "Big data: Transforming the design philosophy of future internet," *IEEE Network Magazine*, vol. 28, pp. 14–19, July/August 2014.

[23] X. Yi, F. Liu, J. Liu, and H. Jin, "Building a network highway for big data: Architecture and challenges," *IEEE Network Magazine*, vol. 28, pp. 5–13, July 2014.

[24] F. Chang *et al.*, "Bigtable: A distributed storage system for structured data," *ACM Transactions on Computer Systems (TOCS)*, vol. 26, 2008.

[25] A. Lakshman and P. Malik, "Cassandra: a decentralized structured storage system," *ACM SIGOPS Operating Systems Review*, vol. 44, no. 2, pp. 35–40, 2010.

[26] D. Agrawal *et al.*, "Challenges and opportunities with big data," tech. rep., Purdue University, 2011.

[27] P. Domingos, "A few useful things to know about machine learning," *Communications of the ACM*, vol. 55, no. 10, pp. 78–87, 2012.

[28] A. Kejariwal, "Big data challenges: A program optimization perspective," in *IEEE Second International Conference on Cloud and Green Computing (CGC)*, pp. 702–707, 2012.

[29] D. Jones *et al.*, "Big data challenges for large radio arrays," in *IEEE Aerospace conference*, pp. 1–6, 2012.

[30] M. Franklin, "Making sense of big data with the berkeley data analytics stack," in *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining*, pp. 1–2, 2015.

[31] X. Xiang, Z. Zhou, and X. Wang, "Self-adaptive on demand geographic routing protocols for mobile ad-hoc networks," in *IEEE INFOCOM 2007 - 26th IEEE International Conference on Computer Communications*, pp. 2296–2300, 2008.

[32] A. Katal, M. Wazid, and R. Goudar, "Big data: Issues, challenges, tools and good practices," in *Sixth International Conference on Contemporary Computing (IC3)*, pp. 404–409, 2013.

[33] S. Subhani, "Autonomous control of distributed energy resources via wireless machine-to-machine communication; a survey of big data challenges," in *IEEE 15th International Conference on Environment and Electrical Engineering (EEEIC)*, pp. 1437–1442, 2015.

[34] R. Gore and S. Valsan, "Big data challenges in smart grid iot (wams) deployment," in *IEEE 8th International Conference on Communication Systems and Networks (COMSNETS)*, pp. 1–6, 2016.

[35] K. Khedo, R. Perseedoss, and A. Mungur, "A wireless sensor network air pollution monitoring system," *International Journal on Wireless & Mobile Networks*, vol. 2, no. 2, pp. 31–45, 2010.

[36] J. Valverde *et al.*, "Wireless sensor network for environmental monitoring: Application in a coffee factory," *International Journal of Distributed Sensor Networks*, vol. 8, no. 1, 2012.

[37] M. Othman and K. Shazali, "Wireless sensor network applications: A study in environment monitoring system," in *Procedia Engineering*, vol. 41 of *International S ymposium on R obotics and Intelligent Sensors*, pp. 1204–1210, Elsevier, 2012.

[38] M. Aminian and H. Naji, "A hospital healthcare monitoring system using wireless sensor networks," *Health & Medical Informatics*, vol. 4, no. 2, pp. 1–7, 2013.

[39] H. Chang, N. Zhou, X. Zhao, and Q. Cao, "A new agriculture monitoring system based on wsns," in *IEEE 12th International Conference on Signal Processing (ICSP)*, pp. 1755–1760, IEE, 2014.

[40] L. Haeberle, "Airborne measurement for state-of-the-art acceptance testing and verification of broadcasting sites," presentation, European Annual Meeting of Regional Broadcasting Organizations, 2014.

[41] G. Kasper, "Phantom eye background," presentation, Boeing Inc., 2014.

[42] J. Leng, "Using a UAV to effectively prolong wireless sensor network lifetime with wireless power transfer," PhD dissertation, University of Nebraska, 2014.

[43] G. Tuna, T. Mumcu, K. Gulez, V. Gungor, and H. Erturk, "Unmanned aerial vehicle-aided wireless sensor network deployment system for post-disaster monitoring," in *In Springer Emerging Intelligent Computing Technology and Applications*, pp. 298–305, 2012.

[44] S. Kamath and J. Lindh, "Measuring bluetooth low energy power consumption," Application Note, Texas Instruments, 2012.

[45] P. Lassila, E. Hyytia, and H. Koskinen, "Connectivity properties of random waypoint mobility model for ad hoc networks," in *In IFIP International Federation for Information Processing*, pp. 159–168, 2006.

[46] L. Lazos and R. Poovendran, "Stochastic coverage in heterogeneous sensor networks," *ACM Trans. Sensor Netw.*, vol. 2, no. 3, pp. 325–358, 2006.

[47] J. Bonnet, P. Subsoontorn, and D. Endy, "Rewritable digital data storage in live cells via engineered control of recombination directionality," in *In USA Nat. Acad. of Sciences*, pp. 8884–8889, April 2012.

[48] D. Hanahan, "Studies on transformation of escherichia coli with plasmids," *J. of Mol. Biology*, vol. 166, pp. 557–580, 1983.

[49] Z. Wang, M. Kim, and G. Rosen, "Validating models of bacterial chemotaxis by simulating the random motility coefficient," in *In Proc. IEEE BIBE*, pp. 1–5, 2008.

[50] W. Guopeng, P. Bogdan, and R. Marculescu, "Efficient modeling and simulation of bacteria-based nanonetworks with BNSim," *IEEE JSAC*, vol. 31, pp. 868–878, Dec. 2013.

[51] M. Gregori and I. Akyildiz, "A new nanonetwork architecture using flagellated bacteria and catalytic nanomotors," *IEEE JSAC*, vol. 28, pp. 612–619, 2010.

[52] L. Cobo and I. Akyildiz, "Bacteria-based communication in nanonetworks," *Els. Nanocomm. Netw.*, vol. 1, pp. 244–256, Dec. 2010.

[53] M. Achtman, N. Kennedy, and R. Skurray, "Cell–cell interactions in conjugating escherichia coli: role of trat protein in surface exclusion," in *In USA Nat. Acad. of Sciences*, pp. 5104–5108, Nov. 1977.

[54] R. Fernandez-Lopez *et al.*, "Unsaturated fatty acids are inhibitors of bacterial conjugation," *Microbiology*, Nov. 2005.

[55] S. Singal, "Radio wave propagation and acoustic sounding," *Atmospheric Research*, vol. 20, no. 2–4, pp. 235–256, 1986.

[56] S. Chiu, D. Stoyan, W. Kendall, and J. Mecke, *Stochastic geometry and its applications.* Wiley, 2013.

[57] J. Teichmann, F. Ballani, and van den Boogaart K., "Generalizations of Matern's hard-core point processes," *Spatial Statistics*, vol. 3, pp. 33–53, 2013.

[58] K. Radnosrati and Y. Koucheryavy, "Trade-offs between compression, energy and quality of video streaming applications in wireless networks.," in *Proc. IEEE ICC*, pp. 1100–1005, 2014.

[59] "Cisco Visual Networking Index: Forecast and Methodology, 2015-2020," white paper, Cisco, 2015.

[60] K. Button, ed., *Infrared and Millimeter Waves V14: Millimeter Components and Techniques, Part 5.* Elsevier, 1986.

[61] S. Rey, "TERAPAN: Ultra-high data rate transmission with steerable antennas at 300 GHz," IEEE 802.15-15-0167-02-0thz, IEEE, Mar. 2015.

[62] J. Boyd, "Fujitsu makes a THz receiver small enough for a smartphone." `http://www.spectrum.ieee.org/tech-talk/telecom/wireless/fujitsu-makes-a-terahertz-receiver-small-enough-for-a-smartphone`, Oct. 2015.

[63] I.-F. Akyildiz, J. M. Jornet, and C. Han, "TeraNets: ultra-broadband communication networks in the terahertz band," vol. 21, pp. 130–135, Aug. 2014.

[64] I. Akyildiz, J. Jornet, and C. Han, "Terahertz band: next frontier for wireless communications," *Physical Communication*, vol. 12, pp. 16–32, 2014.

[65] R. Weiler *et al.*, "Enabling 5g backhaul and access with millimeter-waves," in *European Conference on Networks and Communications (EuCNC)*, pp. 1–5, IEEE, June 2014.

[66] "Channel models for 60 ghz wlan systems," ieee doc. 802.11-09/0334r8, IEEE, 2010.

[67] W. Roh *et al.*, "Millimeter-wave beamforming as an enabling technology for 5g cellular communications: theoretical feasibility and prototype results february 2014.," *IEEE Communications Magazine*, vol. 52, no. 2, pp. 106–113, 2014.

[68] M. Akdeniz, Y. Liu, M. Samimi, S. Sun, S. Rangan, T. Rappaport, and E. Erkip, "Millimeter wave channel modelingand cellular capacity evaluation," *IEEE Journal on Selected Areas in Communications*, vol. 32, pp. 1164–1179, June 2014.

[69] M. Komarov, D. Moltchanov, and Y. Koucheryavy, "Per-source packet-level performance analysis of broadband wireless backhauls carrying traffic aggregates," *Computer Networks*, 2016. In review.

[70] B. Sklar, *Digital Communications: Fundamentals and Applications.* Communications Engineering and Emerging Technologies., Prentice Hall, 2nd ed., 2001.

[71] S. Bose, *An introduction to queueing systems.* Kluwer, 2002.

[72] H. Takagi, *Discrete-time Systems*, vol. 3. North-Holland, 1993.

Summary of Publications

The second part of this thesis includes *six* publications referred to as [P1]-[P6]. None of these publications have been used as part of any other thesis. Works [P3] and [P6] are articles published in scientific journals and the rest are conference papers.

The major contribution of each of the *main* publications is clarified below.

Description of Publications

- **[P1]** M. Komarov, "Network challenges of new sources of big data" , *In Proc. of the 17th IEEE Conference on Business Informatics (CBI 2015)*, 2015. Ch. 1. P. 27-36.

  Description

  The aim of [P1] was to identify and characterize networks as potential sources of Big Data. In addition to classic modern data collection systems, such as WSNs, we highlighted that the nano- and micro-scale networks of the future may impose critical mass of traffic that can be classified as Big Data. We then proceeded describing potential pitfalls in Big Data collection techniques and delivery processes to the handling point identifying these two as the critical challenges of the Big Data generation networks. The next three publications tackle the issues of data generation while the latter two address the problems of information delivery.

- **[P2]** M. Komarov, D. Moltchanov, "System design and analysis of UAV-assisted BLE wireless sensor systems" in *Wired/Wireless Internet Communications, Lecture Notes in Computer Science*, pp. 284–296, Springer International Publishing, 2016.

  Description

  We have started considering the data generation and collection process in Big Data generation networks with conventional WSN systems in [P2]. Particularly, we first highlighted the routed nature modern WSNs as the dominating factor in energy consumption of nodes affecting the network lifetime and preventing the use of WSNs in many applications, where cost and energy efficiency are the two critical factors of interest. We then proceeded proposing a concept of a single-hop non-routed system, where UAVs are used for data collection and then demonstrated that the network lifetime could be drastically increased. We have also argued that once the problem of short network lifetime is solved the application of WSNs may

significantly widen with a lot more data coming from such systems in the near future.

- **[P3]** M. Komarov, B. Deng, V. Petrov, D. Moltchanov, "Performance analysis of simultaneous communications in bacterial nanonetworks" in *Nano Communication Networks*, vol.8, pp.55-67, 2016.

  Description

  We have continued analyzing the process of data collection in potential network sources of Big Data in [P3], where we concentrated on the information delivery in bio-inspired nano/micro systems. Although such systems are characterized by rather limited amount of actual data transmitted, the DNA encoding techniques leads to the large arrays of raw data collected, thus, allowing to classify these systems as potential sources of Big Data. We have used a mathematical apparatus of the absorbing Markov chains to analytically model the process of information delivery in absence and presence of concurrent transmissions.

- **[P4]** A. Nguyen, M. Komarov, D. Moltchanov, "Coverage and Network Requirements of a Flash Crowd Monitoring System Using Users' Devices" in *Lecture Notes in Computer Science,* Springer International Publishing, 2016.

  Description

  The [P4] concludes our investigation of the network sources of Big Data. Here, we have proposed a system for real-time monitoring of instantaneous people gatherings, also known as "flash crowds". We have proposed the principle of using end users' mobile equipment such as smartphones and/or tablets for collecting aural and visual information. The carried out performance evaluation using the developed simulation framework has demonstrated that the modern cellular systems are sufficient to handle aural information load from the monitoring system. For real-time transmission of visual information novel wireless technologies such as those operating in millimeter wave or terahertz frequency bands are needed.

- **[P5]** V. Petrov, M. Komarov, D. Moltchanov, J. M. Jornet, and Y. Koucheryavy, "Interference Analysis of EHF/THF Communications Systems with Blocking and Directional Antennas", *In Proc. of the 2016 IEEE Global Communications Conference (GLOBECOM)*, 2016

  Description

  The problem of information delivery from the network sources of Big Data is investigated in [P5] and [P6]. First, in [P5], the attention is paid to characterizing the mean interference structure in millimeter wave and terahertz systems that can potentially serve as backhauls for Big Data. Particularly, the mean interference as a function of molecular absorption coefficient, antenna directivity angle and different antenna patterns has been investigated. The effect of blocking of high frequency radiation has also been taken into account. In this publication we have shown that for the considered bands the interference may, in fact, be negligible leading to noise-limited wireless channels' performance.

- **[P6]** V. Petrov, M. Komarov, D. Moltchanov, J. M. Jornet, and Y. Koucheryavy, "Interference and SINR in millimeter wave andterahertz communication systems withblocking and directional antennas", *Accepted to IEEE Transactions on Wireless Communications*, 2016.

  Description

  In [P6] we have extended the analysis terahertz networks to the case of signal-to-interference-plus-noise (SINR) metric. We characterized the mean and variance of the SINR as a function of antenna models, absorption coefficient, antenna directivity and blocking. The model has also been extended to take into account the effect of molecular noise. We have first demonstrated that in presence of molecular absorption in the path loss model there is no analytical Laplace transmform inverse of the interference density from a single node and proposed to approximate moments of interference using the Taylor series expansion. We have demonstrated that although the use of highly directive antennas greatly increases the aggregate interference the ultimate effect on SINR is positive. Blocking on LoS paths between interferers and the receiver of interest also increases the mean SINR values. Finally, the type of the antenna model highly affects the results not only quantitatively but qualitatively as well.

Author's Contribution

The research work summarized in this thesis has been mostly carried out in the Department of Electronics and Communications Engineering, Tampere University of Technology, Finland. The author of this thesis is the main contributor to [P1], [P2], [P3]. The author's contribution to [P4],[P5] and [P6] is equal to the contribution of the first authors of the corresponding publications. The reported research has been done by the author, guided by his supervisor Prof. Yevgeni Koucheryavy and by his instructor Dr. Dmitri Moltchanov.

In [P1], the author has worked closely with his instructor defining the problems associated with networks acting as sources of Big Data. Identification of the future nano/micro networking systems as potential source of Big Data is due to the author. The writing load has been done by the author as well.

In [P2], [P3] the author formulated the general problem associated with information delivery in modern routed WSNs and bio-inspired micro/nano networks, respectively. He also specified the models, proposed the modeling approaches and performed analysis of the systems. The author was the principal writer for [P2] and [P3].

The proposed design for flash crowds monitoring system in [P4] was the joint idea of the author and Mr. A. Nguyen. In this publication, the author was also responsible for developing simulation environment and obtaining the associated numerical results. The writing load was fairly divided between the author and Mr. A. Nguyen.

The work in [P5] is a joint study with PhD student Mr. V. Petrov. The author contributed to the system model and was responsible for mean interference calculation applying the Campbell's theorem. The writing load was divided between the author and Mr. V. Petrov.

Similarly to [P5], the work in [P6] has been conducted in close collaboration

with PhD student Mr. V. Petrov. The author performed to derivation of moments of SINR using the Taylor series expansion and computation of first and second moment of interference from a single node using Campbell's theorem. The writing load was fairly shared between the author and Mr. V. Petrov.

Publications

Publication 1

M. Komarov , "Network challenges of new sources of big data" , in *Proc. of the 17th IEEE Conference on Business Informatics (CBI 2015)*, 2015. Ch. 1. P. 27-36.

Publication 2

M. Komarov, D. Moltchanov "System design and analysis of UAV-assisted BLE wireless sensor systems" in *Wired/Wireless Internet Communications, Lecture Notes in Computer Science, pp. 284-296, Springer International Publishing*, 2016.
with permission of Springer.

Publication 3

M. Komarov, B. Deng, V. Petrov, D. Moltchanov, "Performance analysis of simultaneous communications in bacterial nanonetworks" in *Nano Communication Networks, vol.8, pp.55-67,* 2016.

Publication 4

An Nguyen, Mikhail Komarov, Dmitri Moltchanov, "Coverage and Network Requirements of a Flash Crowd Monitoring System Using Users' Devices" in *Lecture Notes in Computer Science, Springer International Publishing*, 2016 with permission of Springer.

Publication 5

V. Petrov, M. Komarov, D. Moltchanov, J. M. Jornet, and Y. Koucheryavy, "Interference Analysis of EHF/THF Communications Systems with Blocking and Directional Antennas", *In Proc. of the 2016 IEEE Global Communications Conference (GLOBECOM)*, 2016.

Publication 6

V. Petrov, M. Komarov, D. Moltchanov, J. M. Jornet, and Y. Koucheryavy, "Interference and SINR in Millimeter Wave and Terahertz Communication Systems with Blocking and Directional Antennas", *IEEE Transactions on Wireless Communications*, 2016.

Tampereen teknillinen yliopisto
PL 527
33101 Tampere

Tampere University of Technology
P.O.B. 527
FI-33101 Tampere, Finland

ISBN 978-952-15-3865-0
ISSN 1459-2045