Energy Efficient and Resilient Internet of Things Networks

IDA SYAFIZA BINTI MD ISA

Submitted in accordance with the requirements for the degree

of

Doctor of Philosophy

The University of Leeds

School of Electronic and Electrical Engineering

October 2019

The candidate confirms that the work submitted is his own, except where work which has formed part of jointly-authored publications has been included. The contribution of the candidate and the other authors to this work has been explicitly indicated below. The candidate confirms that appropriate credit has been given within the thesis where reference has been made to the work of others.

Chapter 4 is based on the work from:

I. S. M. Isa, M. O. I. Musa, T. E. H. El-Gorashi, A. Q. Lawey and J. M. H. Elmirghani, “Energy Efficiency of Fog Computing Health Monitoring Applications,” IEEE Conference on Transparent Optical Networks (ICTON), Bucharest, Romania, 2018.

This paper has been published jointly with my PhD supervisor Prof. Jaafar Elmirghani, my co-supervisor Dr Taisir Elgorashi, Mohamed Musa and Ahmed Lawey from University of Leeds.

Chapter 6 is based on the work from:

I. S. M. Isa, M. O. I. Musa, T. E. H. El-Gorashi and J. M. H. Elmirghani, “Energy Efficient and Resilient Infrastructure for Fog Computing Health Monitoring Applications,” IEEE Conference on Transparent Optical Networks (ICTON), Angers, France, 2019.

This paper has been published jointly with my PhD supervisor Prof. Jaafar Elmirghani, my co-supervisor Dr Taisir Elgorash and Mohamed Musa from University of Leeds.

Chapter 4 and chapter 5 are based on the work from:

I. S. M. Isa, M. O. I. Musa, T. E. H. El-Gorashi and J. M. H. Elmirghani, “Optimized Cloud Fog Energy Efficient Healthcare Monitoring Infrastructure,” to be submitted to IEEE Access.

This paper will be published jointly with my PhD supervisor Prof. Jaafar Elmirghani, my co-supervisor Dr Taisir Elgorashi and Mohamed Musa from University of Leeds.

Chapter 6 is based on the work from:

I. S. M. Isa, M. O. I. Musa, T. E. H. El-Gorashi and J. M. H. Elmirghani, “Resilient Energy Efficient Healthcare Monitoring Infrastructure with Server and Network Protection,” to be submitted to IEEE Access.

This paper will be published jointly with my PhD supervisor Prof. Jaafar Elmirghani, my co-supervisor Dr Taisir Elgorashi and Mohamed Musa from University of Leeds.

Nur Umairah Athirah, Muhammad Nur Irfan Azyze and Ayra Adira,

this PhD thesis is dedicated to
you.

Acknowledgements

In the name of Allah, the Most Gracious and the Most Merciful

All praise to Allah and His blessing for the completion of this thesis. I thank God for giving me the strength, knowledge, ability and opportunity to undertake this research study which requires untiring efforts to persevere and complete it satisfactorily. My humble gratitude to the Holy Prophet Muhammad (Peace be upon him) whose way of life has been continuous guidance for me.

It is a matter of utmost pleasure for me to extend my gratitude to my supervisor, Professor Jaafar Elmirghani, for his guidance, patience and most importantly, for providing me with positive encouragement and a warm spirit throughout my entire PhD journey.

I would also like to express the deepest appreciation to my co-supervisor Dr. Taisir Elgorashi for her guidance during my PhD.

Not to forget, I am very thankful to Dr Mohammed Musa for his help and suggestion during my PhD.

I am also highly indebted to the Higher Education of Malaysia and Universiti Teknikal Malaysia Melaka (UTeM) for fully funding my PhD studies at the University of Leeds. Many thanks to them.

Acknowledgement would be incomplete without extending my gratitude to the most significant source of my strength, my family. I owe thanks to a very special person in my life, my husband, Nur Latif Azyze Bin Mohd Shaari Azyze for his continuous and unconditional love, support and understanding during our pursuit of PhD degree that has made the completion of this thesis possible. Thanks to him for always being there during my ups and downs in this journey. I deeply value his contribution and sincerely appreciate his belief in me.

To my children, Nur Umairah Athirah, Muhammad Nur Irfan Azyze and Ayra Adira, thank you for abiding my ignorance and patience that they have shown during my challenging time; they always assured me. Words would never be enough to say how grateful I am to all of you.

I am forever grateful and would like to extend my heartfelt thanks to my entire family back home in Malaysia for their love and moral support. I thank my mother, Jamaliah Binti Ibrahim, whose dreams for me that have resulted in this achievement and without her loving upbringing and nurturing, I would not have been where I am today and what I am today.

Ida Syafiza

Abstract

Advancement in Internet-of-Things (IoT), mobile technologies and cloud computing services have inspired numerous designs for cloud-based real-time health monitoring systems. However, the massive transfer of health-related data to cloud contributes to increase the congestion in the networking infrastructure which leads to high latency and increased power consumption. Therefore, fog computing is introduced to provide service provisioning close to users. Nevertheless, the energy consumption of both transport network and processing infrastructures have yet to be probed further. Hence, this study proposes a new fog computing architecture under Gigabit Passive Optical Network (GPON) access network for health monitoring applications.

A Mixed integer linear programming (MILP) model is introduced to optimise the number and locations of the processing servers at the network edge for energy-efficient fog computing. The model is developed for GPON and Ethernet access networks used to support fog processing. The impact of equipment idle power and the traffic volume have been investigated, and their effect on energy efficiency to serve low and high data rate health monitoring applications is established. The work also proposes resilient fog processing architectures for health monitoring applications. A MILP model for energy-efficient and resilient fog computing infrastructure considering two types of server protections related to geographic locations of primary and secondary processing servers are developed to optimise the number and locations of the processing servers at the network edge. In addition, a MILP model is developed to optimise energy efficiency and resilience of the proposed fog processing architectures considering server protection with geographical constraints and network protection with link and node disjoint resilience. The impact of increasing the level of resilience on the energy consumption of networking and processing is studied in contexts where the goal is to serve low and high data rate health monitoring applications.

Table of Contents

List of Abbreviations

ABAC	Attribute-based Access Control
AF	Atrial Fibrillation
AGS	Aggregation Switch
AHA	American Heart Association
AHMS	Autonomic Healthcare Management System
APIs	Application Program Interfaces
AQI	Air Quality Index
ASW	Access Switch
AZSPM	Autonomic Zero-Knowledge Security Provisioning Model
BBN	Bayesian Belief Network
BS	Base Station
CA	Conventional Approach
CABG	Coronary Artery Bypass Surgery
CAPEX	Capital Expenditure
CAS	Centre Aggregation Switch
CP-ABE	Ciphertext-policy Attribute-based Encryption
CVD	Cardiovascular Disease
DDoS	Distributed Denial of Service Attack

DMBD	Decoy Medical Big Data
ECG	Electrocardiogram
EECC	Energy Efficient Cloud Computing
EEFC	Energy-efficient Fog Computing
E-HAMC	Emergency Help Alert Mobile Cloud
EOFC	Energy Optimised Fog Computing
EORIG	Energy Optimised Resilient Fog Computing Infrastructure with Geographical Constraints
EORIGN	Energy Optimised Resilient Fog Computing Infrastructure with Geographical Constraints and Link and Node Disjoint
EORIWG	Energy Optimised Resilient Fog Computing Infrastructure without Geographical Constraints
FAAL	Fog Ambient Assisted Living
FOA	Fog Optimised Approach
FPGA	Field-Programmable Gate Array
GeSI	Global e-Sustainability Initiative
GPON	Gigabit Passive Optical Network
H3IoT	Home Health Hub Internet of Things
HL7	Health Level Seven
IAL	Internet Application Layer
ICT	Information and Communications Technology

ILP	Information Processing Layer
IoT	Internet of Things
LCL	Local Communication Layer
LTE	Long-Term Evaluation
LTE-M	Long-Term Evolution for Machine
M2M	Machine-to-Machine
MCC	Mobile Cloud Computing
MCIs	Micro Computing Instances
m-health	Mobile Health
MILP	Mixed Integer Linear Programming
MIT	Massachusetts Institute of Technology
NHS	National Health Services
OLT	Optical Line Terminal
OMBD	Original Medical Big Data
ONS	Office for National Statistics
ONU	Optical Network Unit
OPEX	Operating Expenses
OSA	Obstructive Sleep Apnea
PCE	Power Conversion Efficiency
PCIs	Percutaneous Coronary Interventions Surgery

PFHD	Privacy-preserving Fog-assisted Information Sharing Scheme for Health Big Data
PRB	Physical Resource Block
PS	Processing Server
PSL	Physical Sensing Layer
PUE	Power Usage Effectiveness
QoS	Quality-of-Service
QPSK	Quadrature Phase Shift Keying
RB	Resource Block
RE	Resource Element
RFID	Radio Frequency Identification
SF2CA	Smart F2C Adaptor
SoA-Fog	Service-Oriented Architecture-Fog
SSL	Secure Sockets Layer
TTI	Transmission Time Interval
UAL	User Application Layer
UFW	Uncomplicated Firewall
VM	Virtual Machine
WSN	Wireless Sensor Network

List of Tables

List of Figures

Chapter 1

Introduction

Internet power consumption has continued to increase over the past decade because of the increased traffic of the Internet [1]. Moreover, as a result of the growing popularity of traffic intensive applications, such as Internet Protocol TV and high definition TV, the traffic is expected to grow by about 40% annually for the foreseeable future [2]. As stated in [1], the power consumption of the network is a significant contributor to the total energy demand in many developing countries. For example, in 2005, more than 2 TWh that is approximately 1% of the energy demand in Italy, was used by the Telecom Italia network [1], with similar trends in other developed countries. Meanwhile, British Telecom became the most significant single power consumer in the UK during the winter of 2007, consuming about 0.7% of the UK's total power consumption [1]. Increasing energy efficiency is becoming one of the main priorities for information and communications technology (ICT) organisations, given the ecological and economic drivers that are currently high profile. In view of this, and with the emission of 228 gm of CO_2, whenever the network components consume 1 kWh of traditional electrical energy [3], CO_2 pollution remains high on the agenda.

Moreover, the Global e-Sustainability Initiative (GeSI) reported the expected carbon footprint of networks and related infrastructure at approximately 320 Mtons of CO_2 emissions in 2020 [4], [5]. Mobile communication infrastructure is expected to produce more than 50% of the network CO_2 emission, where telco devices and broadband access equipment will constitute a non-negligible contribution of 22% and 15%, respectively [4]. Because of this, reducing the power consumption while increasing the uptake of services, such as Internet of Things (IoT) based services is very challenging.

The recent increase in chronic diseases and the ageing population have accelerated the developments in remote health monitoring in developed countries [6]. It has been reported that the increasing ageing population over 60 years old will grow from 841 million in 2013 to more than 2 billion in 2050 [7]. The shortage in healthcare professionals also increases the need for an effective health monitoring system that can provide an accurate diagnosis of health data and appropriate treatment. Additionally, the cost of hospitalisation is one of the main factors that have driven the importance of remote health monitoring applications. Previous studies have shown that the conservative estimation of healthcare cost related to Atrial Fibrillation (AF) by the UK's National Health Services (NHS) was between £244 million and £531 million in 1995 and the amount doubled in 2000 [8]. The additional cost due to the admission of patients, where AF was the primary diagnosis, is £221 million while hospitalisation due to heart failure or stroke, which is treated as the second position, is £228 million [8]. Moreover, the cost for post-discharge outpatient visits and the cost of nursing home care related to the AF is £31.7 million and £110.7 million, respectively [8].

Therefore, the implementation of remote health monitoring with patient-centric healthcare, where the hospital, patient and services are seamlessly connected, is an effective way to provide the service to the patient while reducing the operational costs. Also, it has been reported, [9] that the delivery model of healthcare from the present hospital-centric system will be transformed to hospital-home-balance by 2020 and to a final home-centric system by 2030.

The advancement in wireless body sensors and mobile technologies has motivated the progression of the mobile-based health monitoring system (m-health). The introduction of m-health services can provide real-time feedback to the patient about their health condition, as well as alerts on health-threatening conditions. Also, the rapid growth in cloud computing has enabled the development of mobile cloud computing (MCC) applications that offer high processing and storage capabilities for health data. In fact, exploiting various disciplines within the health arena using machine learning methods to perform early-detection and prediction of diseases is indeed an effective strategy towards enhancing healthcare systems [10].

Nonetheless, the massive transfer of health-related data from patients to the cloud contributes towards increasing the congestion in cloud networking infrastructure. Hence leading to high latency and potential violations of Quality-of-Service (QoS) metrics [11]. These also increase the occurrence of errors where the impact of a single error in the analysed data can cause inaccurate treatment decisions, especially for emergency cases, which can be critical [12]. Furthermore, the large volumes of transferred data can increase the energy consumption within the network as the data have to travel over the network to reach the cloud for processing [13].

One effective way to address this limitation in cloud networks is by bringing service provisioning closer to users while maintaining mobility among them

[14]. However, as both locations and deployments of data centres are fixed, a new paradigm, which is referred to as ‘fog computing’, has been introduced as a new platform by Cisco, to overcome this shortcoming [15]. Fog computing is often referred to as edge computing [16]–[21]. However, the Open Fog consortium [22] has clearly expressed that fog computing differs from edge computing where fog works with the cloud while edge computing is an exclusion of cloud [23]. Fog computing extends the cloud-based Internet by initiating an intermediate processing layer using fog servers for example, between IoT devices and the cloud. The fog servers are a highly virtualised computing system equipped with data storage, as well as computing and communication facilities, which appear similar to the cloud servers [18]. It is also possible to connect the fog servers to the cloud to leverage the rich functionality and available application tools. Furthermore, as fog servers are disseminated at the network edge, fog computing may have dense geographical coverage and mobility support. Therefore, fog computing can deliver QoS metrics in healthcare monitoring systems for patients due to reduced latency besides reducing the energy consumption in cloud networking infrastructures [24]–[32]. Recent studies have applied fog computing to develop efficient health monitoring systems. For example, a monitoring system [12] employed the concept of fog computing at a smart gateway to efficiently process health data, particularly the electrocardiogram (ECG) signal. The ECG empirical results for feature extraction using the proposed system displayed 90% bandwidth efficiency and low latency, real-time response. Additionally, the authors of [10] claimed that both continuous monitoring and real-time monitoring might be dysfunctional with the present IoT-based systems. Therefore, fog computing was embedded in the system

and the results exhibited reduced response time and increased system reliability in the presence of intermittent Internet connections. A prototype of a smart e-health gateway [33] (i.e. a fog computing device) has been implemented to reduce the burden at the sensor node and the cloud by performing high-level services such as real-time data processing, local storage and embedded data mining. The gateway performance is evaluated in terms of the energy efficiency of the sensor nodes, scalability, mobility and reliability. Meanwhile, a real-time event triggering for health monitoring systems in smart homes [27] was proposed by implementing a Bayesian Belief Network (BBN) to classify the state of events at the fog layer.

Despite the effectiveness of fog computing functions in health monitoring applications in terms of delivering real-time monitoring with lower latency and lower bandwidth utilisation, most studies have dismissed the essential aspect of the energy consumption in transport networks. In this study, a new framework for an energy-efficient health monitoring system that performs real-time monitoring and a patient-centred environment are considered leveraging the concept of fog computing. Fog computing has been identified as a potential paradigm that can contribute to reducing the energy consumption of networking infrastructure and processing while providing the same health monitoring services as the cloud computing for energy efficiency purposes. The West Leeds area was considered as a case study to examine the energy efficiency of fog computing for health monitoring applications. Mixed integer linear programming (MILP) models and real-time heuristic algorithms have been developed to model and solve the problems of optimum network and processing resource allocation with the goals of reducing the energy consumption of networking equipment and processing for ECG monitoring and video fall monitoring applications. Also, a further MILP model and a further real-time heuristic are developed to increase the energy-efficiency of a resilient version of the infrastructure used in the proposed health monitoring applications. The resilience was improved in this IoT-fog-cloud architecture by considering server and network protection under both ECG and video fall monitoring applications.

1.1 Research objectives

So far, the research reported in the literature has considered the use of fog computing for health monitoring applications. It did not consider the energy consumption of the network used in transporting the raw health data to the fog for processing and transporting the analysed health data from the fog for feedback and from the fog to the cloud for permanent storage. Also, the impact of the fog locations in the edge network on energy efficiency has not been studied. The hypothesis in this study is that optimising the locations of the fog at the network edge can improve the energy efficiency of the fog-based networking infrastructure and processing for health monitoring applications. In this research, therefore, the fundamental objectives are as follows:

1. Propose a fog computing architecture for health monitoring applications using a Gigabit Passive Optical Network (GPON) access network. The candidate locations of the fog are intended to be only at the access layer.
2. Investigate and improve the energy efficiency of fog computing by optimising the number and locations of fog servers at the access layer to serve low and high data rate health monitoring applications.

3. Investigate the energy efficiency of resilient fog computing infrastructure for server protection with and without geographical constraints by optimising the number and locations of primary and secondary processing servers at the access layer to serve ECG and video fall monitoring applications, separately. Here the geographic constraints improve resilience by preventing the co-existence of the primary and secondary servers at the same node.
4. Investigate the energy efficiency when network protection is added under link and node disjoint resilience with resilient fog computing infrastructure for server protection with geographical constraints to serve ECG and video fall monitoring applications, separately.

1.2 Original contributions

1. Designed a fog computing architecture using a GPON access network and proposed optimum the locations to place the fog at the access layer to minimise power consumption.
2. A novel MILP model was developed to optimise the number and locations of the processing servers at the access layer so that the energy consumption of both networking equipment and processing are minimised. Also, a real-time energy optimised fog computing (EOFC) algorithm was developed as a real-time implementation of the proposed MILP model.
3. Developed a second MILP model to optimise the number and locations of primary and secondary processing servers at the access layer considering server protection without geographical constraints for

energy efficiency. Also, a real-time Energy Optimised Resilient Fog Computing Infrastructure without Geographical Constraints (EORIWG) algorithm was developed as a real-time implementation of the proposed MILP model.

4. Developed a third MILP model to optimise the number and locations of primary and secondary processing servers at the access layer considering server protection with geographical constraints for energy efficiency. Also, a real-time Energy Optimised Resilient Fog Computing Infrastructure with Geographical Constraints (EORIG) algorithm was developed as a real-time implementation of the proposed MILP model.
5. Developed a fourth MILP model to optimise network protection with link and node disjoint resilience for energy efficiency while also optimising the number and locations of primary and secondary processing servers at the access layer considering server protection with geographical constraints. Furthermore, a real-time Energy Optimised Resilient Fog Computing Infrastructure with Link and Node Disjoint Constraints (EORIN) algorithm has been developed as a real-time implementation of the proposed MILP model.

1.3 Related publications

The work in this thesis resulted in the following papers:

1. I. S. M. Isa, M. O. I. Musa, T. E. H. El-Gorashi, A. Q. Lawey and J. M. H. Elmirghani, “Energy Efficiency of Fog Computing Health Monitoring Applications,” IEEE Conference on Transparent Optical Networks (ICTON), Bucharest, Romania, 2018.

2. I. S. M. Isa, M. O. I. Musa, T. E. H. El-Gorashi and J. M. H. Elmirghani, “Energy Efficient and Resilient Infrastructure for Fog Computing Health Monitoring Applications,” IEEE Conference on Transparent Optical Networks (ICTON), Angers, France, 2019.
3. I. S. M. Isa, M. O. I. Musa, T. E. H. El-Gorashi and J. M. H. Elmirghani, “Optimized Cloud Fog Energy Efficient Healthcare Monitoring Infrastructure,” to be submitted to IEEE Access.
4. I. S. M. Isa, M. O. I. Musa, T. E. H. El-Gorashi and J. M. H. Elmirghani, “Resilient Energy Efficient Healthcare Monitoring Infrastructure with Server and Network Protection,” to be submitted to IEEE Access.

1.4 Thesis organisation

Following the introduction in this chapter, this thesis is organised as follows:

Chapter 2 reviews the concept of the Internet of Things (IoT) and the applicability of IoT in remote healthcare applications. An overview of cloud-driven and fog-driven IoT healthcare applications is presented. Also, various approaches to develop IoT remote healthcare systems with fog computing are described.

Chapter 3 introduces a novel fog computing architecture under a GPON access network for health monitoring applications and introduces the system flow of the health monitoring applications under fog and cloud architecture. A MILP model is introduced to optimise the number and locations of the processing servers at the access layer for energy-efficient fog computing (EEFC). Also, a MILP model for the conventional approach (i.e. energy-efficient cloud computing, EECC) is developed as a benchmark to evaluate the performance of the proposed fog approach in terms of the networking equipment and processing energy consumption.

Chapter 4 introduces the methodologies used to determine the model input parameters for ECG monitoring applications to evaluate the performance of the proposed fog computing (EEFC model) and cloud computing (EECC model). The energy-optimised fog computing (EOFC) heuristic algorithm is also developed based on the insight from the results in the EEFC model for real-time implementation. The EEFC model, EECC model and the EOFC heuristic algorithm are tested under two kinds of access network; GPON and Ethernet. Also, the effect of the equipment idle power and the traffic volume are evaluated for the EEFC model and for the EECC model and are compared.

Chapter 5 introduces the parameters considered for the video fall monitoring application to evaluate the performance of the EEFC and the improvements over the EECC. The two models developed for EEFC and EECC are tested under two scenarios: (i) limited number of patients that can be served in a single processing server; and (ii) limited number of processing servers that are allowed at each candidate node. In addition, the performance of the EOFC heuristic algorithm under both scenarios is presented and compared with the EEFC model.

Chapter 6 introduces the proposed resilient fog computing architecture for health monitoring applications. The performance of fog computing in terms of energy consumption of both network and processing is investigated considering two types of server protection: with and without geographical constraints. A mixed-integer programming model (MILP) optimisation model was developed for energy-efficient and resilient fog computing infrastructure considering the two types of server protection together with the corresponding heuristic model for real-time implementation. The performance of the fog computing in terms of energy consumption of both network and processing with increasing level of resilience considering server protection with geographical constrains and network protection with link and node disjoint resilience is also investigated. The constraints related to the network protection, in addition to the previous MILP model, are introduced for energy-efficient and resilient infrastructure for fog computing, considering both server protection with geographical constraints and network protection with link and node disjoint resilience. Also, the heuristic model is developed for real-time implementation. The models are tested separately under two different monitoring applications, ECG and fall.

The thesis is concluded in Chapter 7, where the significant contributions of this work are summarised, and future directions are discussed.

Chapter 2

Review of IoT and cloud-based healthcare

2.1 Introduction

In this chapter, a review of the Internet of Things (IoT) and the applicability of the IoT in remote healthcare applications is reviewed. The IoT is expected to make radical changes in a range of industries and in our daily life. The IoT offers a seamless platform to connect people and objects, hence enriching and making our life more comfortable. As the IoT network can consist of billions of sensors, the generation of enormous amounts of data-on-demand is increasing significantly. Therefore, an overview of the cloud and fog-driven IoT architecture, which are the key enablers of the IoT vision, is presented. This chapter also presents an essential review of the proposed architecture of the IoT remote healthcare system which has been created to improve healthcare services.

2.2 Internet of Things networks

The Internet of Things (IoT) was coined by Kevin Ashton, who co-founded the Auto-ID Centre at the Massachusetts Institute of Technology (MIT) in 1999 [34], [35]. He was one of the first people to see the potential of developing a system where the Internet connects to the physical world via ubiquitous sensors [34]. Nowadays, the IoT is an ever-growing ecosystem that integrates physical devices, animals and people over a network, thus enabling them to interact, communicate, collect and share data anytime and anywhere [36]. With the advancement of modern wireless telecommunications, the IoT has become a novel paradigm that is rapidly gaining ground. The main concept of the IoT includes intelligent communication between connected devices with less human intervention [37]. Therefore, a typical IoT system is equipped with sensors, communication interfaces, computational and processing units and cloud interfaces [36], [38]. The sensor is used to collect data from different devices. The purpose of the communication interfaces, such as those used in Radio Frequency Identification (RFID) and Wireless Sensor Network (WSN) technologies, is to provide the means of communication and networking. The processing and analysis units are designed to process and analyse the data through Application Programme Interfaces (APIs). Moreover, the purpose of the cloud interfaces is to provide access between the devices and the cloud computing, which can execute more complicated algorithms.

The emergence of the IoT and the pervasive connectivity between people and processes would allow services to be automatically delivered whenever and wherever required. The integration of connected smart devices and cloud-based services aids in addressing the pressing issue of energy efficiency and security at home and in the city via remote monitoring [35], [39]. The IoT can also improves transportation systems, as most of the vehicles are equipped with sensors, actuators and processing power to update the status of the cars for safety purposes. Even roads can be fitted with tags (i.e. RFID and sensors) to send relevant information to traffic control sites and vehicles to better route the traffic. The improvement of the transportation system could lead to the development of a smart city. Moreover, the IoT can aid in improving the quality of care for patients, as the connected smart devices will help by supporting a range of e-health services. The wide range of e-health services supported by intelligent devices enhance access and enable the monitoring of chronic diseases and age-related conditions in the home [39].

However, the IoT also poses several challenges for the network and data centres due to the vast number of IoT devices. Cisco predicts that 50 billion devices will be connected to the Internet by 2020 [40] - [41]. Meanwhile, as forecast by GSMA, the number of globally interconnected devices will increase from 9 billion in 2011 to 24 billion in 2020, 50% of which will be mobile [42]. The mobile devices and other physical objects equipped with sensors are mainly used for computation and data access. Therefore, increasing the number of those devices will lead to a rise in the amount of data generated [43]. However, the IoT devices have limited memory, storage and processing capability, as they are designed to be low-cost and light-weight. Therefore, it would be impossible for the data to be fully processed and stored locally. Hence, massive processing and computation resources must be available ideally on per-use basis or rental basis [44].

Cloud computing technologies provide services for computation and storing at anytime and anywhere. Therefore, cloud computing infrastructure may seem the best choice when it comes to deploying IoT platforms [15]. The data which cannot be processed locally can be offloaded to the cloud through the Internet for processing. As the cloud is centralised, using cloud computing provides opportunities for cost-saving [45]. However, offloading massive amounts of data generated by the IoT devices to the cloud for computation increases the congestion in the network. The high congestion can lead to a rise in the latency, which in turn threatens the performance and usability of applications. Moreover, the considerable distance from the IoT devices to the cloud increases the latency, and this is not ideal for certain applications which require low latency and short response times.

Introducing fog computing can provide solutions to the latency problem associated with cloud computing. Fog computing is a new concept introduced by Cisco that extends the cloud computing paradigm to the edge of the networks. Fog provides services in terms of computation and storage, similar to the cloud paradigm. The advantage of fog is its proximity to the users, which means low latency and fast response time, and this suits most of the IoT applications' requirements [46]. In addition, fog can have dense geographic coverage, as fog servers are distributed at the edge of the network, which offers fault tolerance, reliability and scalability of the system [15]. Therefore, in utilising the fog computing platform, IoT applications and services can be operated locally (i.e. edge of network), hence saving bandwidth.

2.3 IoT for healthcare

One of the most attractive application areas for IoT is medical care and healthcare, which gives rise to many medical applications, especially remote health monitoring [47]. The website MarketResearch.com also predicts that the IoT usage for healthcare will reach $117 billion by 2020 [48]. The implementation of remote health monitoring requires a transition from hospital-centric treatment to patient-centric healthcare where the hospital, patient and services are seamlessly connected. It has also been reported that the delivery model of healthcare from the present hospital-centric treatment will be transformed into hospital-home-balance in 2020 and into home-centric in 2030 [9].

The IoT devices often run on low energy. However, there is a restriction on processing and storage capabilities [49]. The limitation of the computation and storage capability of the IoT medical devices and wearable devices to process the health data requires computation offloading. The rich functionality of cloud computing has been widely accredited to support the IoT-enabled healthcare solution. This enables the computation offloading of big data from the medical and wearable devices to the cloud for processing, analysis and storage purposes.

There are many health applications [19] – [21] which have been developed using the central cloud for processing, analysis and storage units. The cloud is used as it offers patients and doctors access to the stored data at anytime and anywhere. However, the centralised location of these cloud data centres requires health data from the medical and wearable devices to travel long distances for processing. This increases the latency in the network, which may endanger the patients. Besides, the cloud also provides services to other applications such as traffic monitoring, surveillance monitoring, etc. This increases the amount of traffic to the cloud, hence leading to a rise in the latency in the network. As health monitoring is a latency-sensitive application, offloading the health data to the cloud for processing is not an effective approach.

Fog computing is a promising solution that offers local processing for health data. Fog computing offers an almost similar function as the cloud, but with limited processing and storage capability. Although fog resources have a limitation when it comes to computing and storage capability, they are flexible enough to be customised according to the application context [53]. In fog computing, the fog servers are distributed at the edge of the network, which is closer to the user, therefore reducing the latency in the network. Furthermore, fog computing minimises the energy consumption in cloud networking infrastructures [23] – [31] under increasing applications traffic.

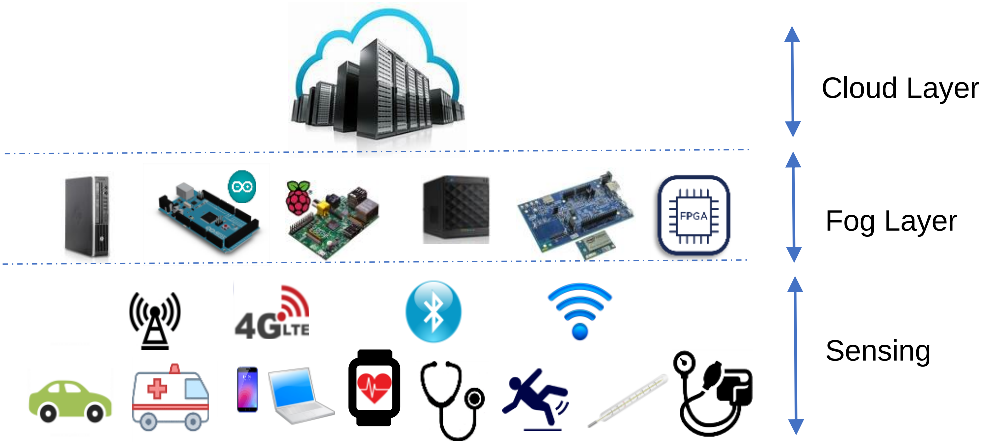


Figure 2.1: Multi-layer architecture of health monitoring applications

Figure 2.1 shows a multi-layer architecture of health monitoring applications which consists of three main layers, i.e. sensing layer, fog layer and cloud layer. The sensing layer consists of health and wearable devices that are used to collect data from patients. The fog layer includes small computing devices that perform local processing and analysis. The health data at the sensing layer is sent to the fog for processing and analysis, through various communication protocols supported by fog devices such as Wi-Fi, ZigBee etc. The cloud layer includes high performance computing capabilities units and is used to perform complex tasks such as long-term data analytics.

The results or data set obtained from this analysis are shared among the fog servers, as the fog and the cloud can communicate through an Internet connection.

2.4 Cloud-based health monitoring applications

Telehealth refers to both clinical and remote non-clinical services that involve the use of telecommunications and virtual technology in order to deliver healthcare outside of the traditional healthcare facilities. The aim of telehealth is to improve the quality and efficiency of healthcare while reducing its cost through the use of electronic means. The authors in [54] highlighted four modes of telehealth activities. The first is store-and-forward-telehealth. In this mode, all of the recorded health data, including images, video and audio, is transmitted to other locations when needed. The second is real-time telehealth data. In this mode, the patients can communicate directly with the doctors through videoconferencing. The third is remote patient monitoring, where patients send their health data to the cloud, for example, and to the doctor who forms a diagnosis based on the data and sends feedback regarding the recommended treatment to the patients. The fourth is remote training, which focuses on patients with chronic diseases, and also focuses on providing sophisticated care to the patients over the network.

Mobile health (m-health) also plays a significant role in supporting healthcare via electronic means. m-health utilises cloud platforms, which allows patients to access their health information through their smartphone apps or web-based cloud dashboard at anytime and anywhere [36]. However, the main difference between telehealth and m-health is that m-health only uses mobile devices or other wireless devices such as mobile phones or tablets. In m-health, mobile devices are used to continuously track and manage specific health data using apps without the need for the person's healthcare provider. However, m-health also allows the patients to communicate with their physicians without meeting face-to-face.

There exist several works focused on developing the m-health system architecture. For example, in [50], an m-health system architecture considering emergency scenarios was proposed, where the mobile or smartphone of the patient is a powerful tool for service access and delivery. The proposed system is based on the smart space paradigm, which adopts IoT technologies and Semantic Web. The system consists of patients, medical personnel, healthcare services and other participants who operate in the same network computing environment, and they interact by sharing the information and its semantics. The proposed system is also useful in the case of patient mobility. Figure 2.2 shows the m-health system architecture that supports personalised assistance services for emergency cases with remote and mobile patients [50].

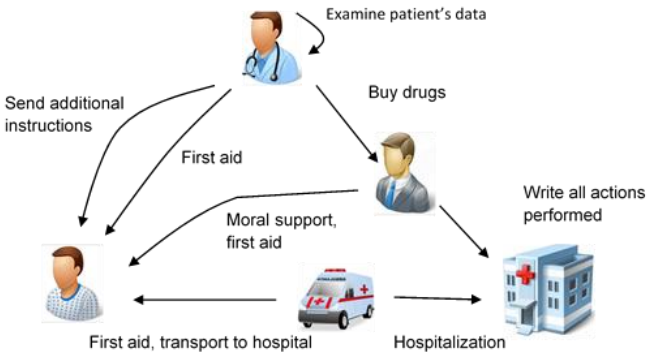

Examine patient's data
Send additional instructions
Buy drugs
First aid
Write all actions performed
Moral support, first aid
First aid, transport to hospital
Hospitalization

Figure 2.2: m-Health system architecture considering emergency scenarios [50]

As reported in [55], there are more than 85 publicly-available Internet sites on diabetes self-management. In [51], an IoT-based platform to support self-management of diabetes was proposed. The work aims to integrate multidimensional aspects of treatment and develop a patient-centred approach. Therefore, to achieve the aims, the smartphone platform is designed to offer remote manageability capabilities and enables care over distance, as found in [50]. Figure 2.3 illustrates various aspects of platform support for diabetes therapy [51]. First is data collection, where the information regarding daily self-management of the disease, measurements from the medical sensors and short messages from patients are sent to the health portal. Second is feedback from the system to the patients where, based on the collected reading, the mobile phone provides the necessary feedback. The third is feedback from caregivers to the patients, where the caregivers provide treatment plans to their patients.

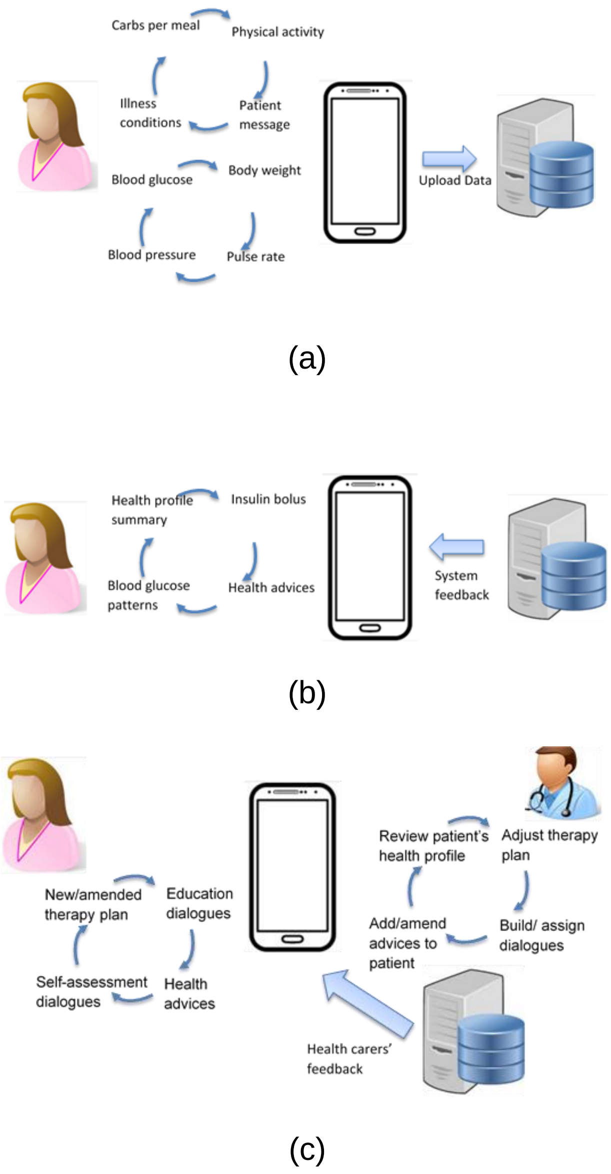


Figure 2.3: Aspects of platform support for diabetes (a) Data collected from the patients (b) System feedback to the patient (c) Caregiver feedback to the patient [51]

Another architecture framework, namely the Home Health Hub Internet of Things (H3IoT) was proposed in [52] to monitor the elderly at home. The architectural concept of H3IoT consists of five layers, as shown in Figure 2.4. The first is the physical sensing layer (PSL), which includes medical sensor devices. The second is the local communication layer (LCL), where the data is transferred to the upper layer using wireless technologies such as Zigbee.

The third is the information processing layer (IPL), where the health data is processed. The fourth is the Internet application layer (IAL), which is the backbone of the system. The IAL performs data storage and analysis at a later stage. The last layer is the user application layer (UAL), where the doctor or hospital receives real-time information about the health status of elderly patients. Moreover, the proposed H3IoT framework provides mobility besides lower cost and delay tolerance.

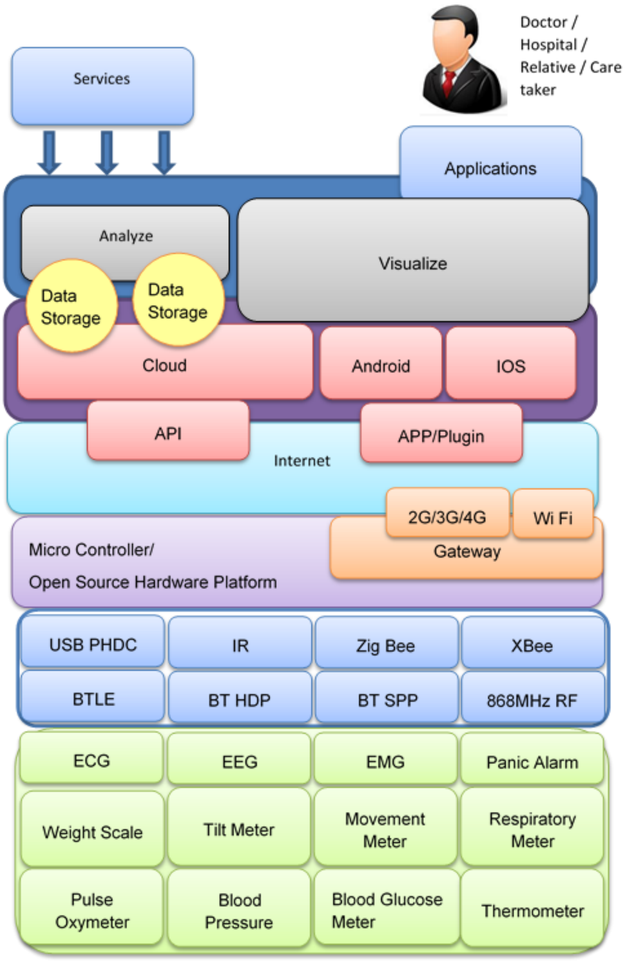


Figure 2.4: H3IoT architecture framework for health monitoring for elderly [52]

2.5 Roles of fog computing and cloud computing in health monitoring applications

The advancement of wireless sensor body network technology is accelerating the development of health monitoring applications [54], [56], [57]. However, the low computation and storage capabilities of the sensor body network and the smart devices have restricted the equipment's ability to process the health data, which requires a sophisticated algorithm. The emergence of cloud computing, which offers high processing and storage capabilities, has contributed to the advancement of healthcare provision [45]. However, the vast amount of health data being transferred to the cloud has resulted in an inevitable increase in the burden on the cloud. Due to this, fog computing has been introduced to allow some tasks to migrate from the cloud to the network edge. With fog computing, health data is processed locally, hence reducing the burden on the cloud. The fog servers are distributed over the edge network to provide high coverage, thus increasing the cost of installations. Moreover, the computation and storage capability of fog servers is far lower compared to the cloud. Therefore, both fog and cloud, with their different roles, are essential in developing health monitoring applications.

2.5.1 Roles of fog computing in health monitoring applications

In this section, the roles of fog computing in healthcare provision are discussed as follows:

2.5.1.1 Switching network

The fog devices receive heterogeneous data from various devices [58]. Therefore, the fog needs to support many protocols such as ZigBee, Bluetooth, Wi-Fi, etc. This is to ensure that the fog devices can read and process the data. However, due to certain factors, the fog may need to transmit the pre-processed data to the cloud for further analysis. As the cloud supports only TCP/IP protocols, the fog needs to activate the network between the IoT and the cloud before sending to the cloud.

2.5.1.2 Data processing

The fog devices are equipped with computation and storage capabilities. Therefore, the fog should have the capabilities to perform local data processing [36] – [39]. This is essential for health applications that require real-time processing. However, as the computation in fog devices is limited, fog should at least process simple tasks by leveraging the data set provided by the cloud. The simple tasks include pre-processing the health data to eliminate noise from signals and to extract useful knowledge for further analysis [59]. This can help to improve the healthcare services by minimising latencies, as the fog servers are located near the users.

2.5.1.3 Pushing services

Fog acts as an intermediate layer between the end devices and the cloud. Therefore, fog needs to provide pushing services to ensure reliable and efficient delivery of the other services. The pushing services provided by fog include receiving the data from the end devices for processing and uploading the processed data to the cloud for further analysis or permanent storage [58].

2.5.1.4 Filtering

Filtering is one of the essential jobs for fog. Some of the health applications require a large amount of health data, such as Electrocardiogram (ECG) signal. As this health data may require a sophisticated algorithm for processing, cloud computing is needed. Transmitting the large amount of health data directly to the cloud will increase the latency. Thus, fog should provide filtering services [36], [37], [39] so that only useful information extracted from the raw signal is sent to the cloud. Moreover, the filtering at the fog servers can eliminate the redundant data, hence reducing the congestion and latency in the network.

2.5.1.5 Aggregation / Data fusion

Data fusion in healthcare is a process that integrates various kinds of health data from numerous devices to produce more accurate and useful information than that provided by an individual data source [37], [39], [61]. Allowing fog to offer data fusion can help reduce the bandwidth utilisation, as only one piece of data will be sent to the cloud for further action. However, the integration of diverse data sources will consume time, thus increasing the processing latencies. This is even worse for applications that involve periodic monitoring. Therefore, as discussed in [35], extending the task at fog to perform data fusion may not be efficient for specific health applications that require periodic monitoring.

.

2.5.1.6 Channel management

Channel management is essential [59] in order to avoid channel conflicts due to aggregating data from the various sensors, which may cause incorrect data to be delivered at the fog receiver. As fog receives heterogeneous data from multiple devices, each sensor node or group of sensor nodes is assigned a specific channel. Moreover, the channel management services verify the incoming data regularly [59]. Therefore, when an abnormality is detected in any channels, request messages will be sent to that channel, following which the services will wait for an acknowledgement message from the sensor nodes. In case of conflict, a push notification is sent to the system administrator to flag up the problem.

2.5.1.7 Data compression

Data compression [37], [39] at fog can reduce the communication latency while increasing the energy efficiency of the network. Two types of data compression which are widely used in the Health-IoT system are lossy and lossless compression. Lossless data compression methods require high processing speed and large memory size, as they perform complex algorithms. Meanwhile, the lossy data compression method requires low processing and low memory, thus meaning it is suitable for resource-constrained sensors.

2.5.1.8 Data storage

The fog devices are equipped with storage capability. Therefore, fog devices store the incoming data in their local storage banks [36] – [38]. This data is stored in a compressed or encrypted way. Moreover, fog is used to temporarily store the analysed data before it is sent to the cloud for permanent storage. Besides, the fog storage can be used as a cache to implement continuous data flows to reduce bandwidth requirements in the metro and core network.

2.5.1.9 Security

Security is one of the main requirements for healthcare applications. Therefore, the privacy of patients’ health data should be guaranteed and the data should be protected from unauthorised accesses [59]. Due to this, fog should apply the security, cryptography and authentication methods at the fog layer [59].

2.5.2 Roles of cloud computing in health monitoring applications

In this section, the roles of cloud computing in healthcare provision are discussed.

2.5.2.1 Data mining

Data mining in healthcare is the process of analysing large data sets to extract useful information which makes it possible to identify meaningful patterns [35],

[58]. Those patterns are used to predict new diseases and assist the doctors in making their clinical decision [62]. Therefore, data mining needs high processing capability, as it requires a sophisticated algorithm. Due to this, data mining is performed in the cloud rather than in fog, as it has limited computing capability [58]. Following this, the extracted information is shared with fog to ensure that the fog server which processes the health data can give proper advice regarding treatment and more so regarding immediate actions to the patients.

2.5.2.2 Permanent storage

Cloud computing offers a high storage capacity. Therefore, the cloud is usually employed to store [42] – [44] information related to the analysed health data for further analysis and history purposes. In addition, the cloud offers data access to authorised users such as patients and doctors.

2.5.2.3 Predict risk stratification

Risk stratification is performed to identify and predict the probability of patients who suffer from any disease [65]. This is important for preventive purposes to avoid the worst outcome, as such an outcome may risk the patient’s life. However, to identify the risk stratification of the patient, a sophisticated algorithm is used, which requires high processing capabilities.

2.6 Fog-assisted health monitoring applications

There exist many pieces of research focused on developing a health monitoring system that leverages fog computing. Each of the studies contributes to different goals for health monitoring applications such as energy, latency, accuracy, security and cost. In this section, works on fog-assisted health monitoring applications are reviewed, and the contribution of the work which has adopted the networking perspective is highlighted.

2.6.1 Energy efficiency

Energy efficiency is one of the primary concerns when developing a health monitoring system to reduce healthcare costs and improve the quality of healthcare services [59]. For instance, in pervasive health monitoring, utilising energy-efficient medical devices for data collection can reduce the chances of system failure, which would risk the patient's life. The energy-inefficient devices may need to have their batteries replaced regularly, which impedes the use of sensors [66].

Energy-efficient and portable sensor nodes for health monitoring applications were developed in [59]. The evolved sensor node was tested to monitor the ECG signal. The results revealed that the sensor nodes consumed less energy in addition to having a long operating time of up to 155 hours. In [67], RF and solar energy harvesting were designed for health monitoring applications to power up the sensor nodes. These sensor nodes were used to collect health data and send it to fog wirelessly via a wireless module. Here, a novel antenna which is able to receive RF power from GSM, Wi-Fi, Bluetooth, 3G and LTE was designed. Moreover, the authors developed a RF-DC rectifier that achieved high power conversion efficiency (PCE) levels at lower RF input power levels.

Meanwhile, the research in [63] investigated the lifetime of the wearable devices used to capture the ECG signal based on the number of channels and sampling rate. The results indicated that the device lifetime increased when low channels and low sampling rate were utilised to record an ECG signal. In addition, the results showed that transferring real-time data to the mobile device consumed more energy compared to transferring non-real-time data. The authors also claimed that applying compression techniques to reduce the amount of data would immediately affect the battery life of the devices.

The type of technology used to design energy-efficient fog devices is very important. In [65], Field-Programmable Gate Array (FPGA) technology, was chosen to design the fog nodes. The FPGA node is reconfigured to produce maximum performance in tasks with low power consumption. Meanwhile, in [16], an experiment was conducted to evaluate the performance of two fog devices, namely Raspberry Pi and Intel Edison in terms of power consumption. The results indicated that the Raspberry Pi consumed less power compared to Intel Edison. In [33], the authors claimed that data processing at sensor nodes will consume high levels of energy. Therefore, a fog gateway, UT-GATE, was proposed to perform data fusion and data compression of health data. The results revealed that 55.7% energy saving was achieved by processing the data at the gateway compared to the sensor node.

The integration of cloud-fog services into the healthcare solution was explored in [22]. The proposed system was evaluated via simulation using the iFogSim simulator. The results revealed that the integration of cloud-fog is essential to overcoming the increasing demands for processing that may require a huge number of sensors or vast CPU requirements to serve the application. Moreover, the results showed that the integration of fog-cloud increased the performance of the fog-based solution in terms of service distribution, instances cost, energy usage and network delay. The authors in [7] studied the relationship between efficient resource management in fog and the energy consumption. They proposed a QoS-aware resource allocation algorithm to optimise the resource utilisation of fog at the edge network before deciding to offload the processing request to the cloud. The results revealed that fog contributed to lower the energy consumption compared to the cloud under an increasing number of demands.

2.6.2 Latency

Fast-response time is one of the primary requirements in health monitoring applications. The delay in detecting abnormal health signals may cause harm to the patients. Therefore, to improve the response time, the latency issues have to be taken seriously. Many works have been widely performed to overcome the latency issues in the health monitoring system. Such works include designing the system architecture of the health monitoring applications, designing the network architecture of the health applications, and optimising the resource utilisation at the fog and cloud.

In [68], [69], the eWALL system was integrated into fog computing for real-time processing. The system aims to reduce the communication overhead by processing the health data at the network edge, therefore reducing the latency in the network. Meanwhile, in [70], the results revealed that the type of classifier used to classify the patient’s health state may also result in high response time. Various kinds of classifier algorithms were tested, including Bayesian Belief Networks (BBN), neural networks, k-nearest neighbour and linear regression. The results showed that the BBN classifier had the highest response time when it came to determining the patient’s condition compared to other algorithms.

Moreover, a real time fall detection system, called U-Fall to detect fall among stroke patients was proposed in [71]. In this work, the analytic tasks were distributed throughout the network; this involved splitting the detection task between the edge devices (i.e. smartphone) and the cloud server for fast response time. In the above work, the light-weight computation for fall detection will be conducted by the edge device. Besides this, the data from the sensor will be sent to the cloud for processing to achieve accurate detection. The results reveal that the response time of the U-Fall system is close to the minimum of the existing two approaches, namely the T-system, which is based on a threshold technique, and the P-System, which is based on a pattern matching system.

An architectural approach to the autonomic healthcare management system (AHMS) was proposed in [72]. The above work aims to provide fast response action to the patient in case of falls. Therefore, a local autonomic manager is embedded in fog to collect information from the wearable devices and to check the occurrence of an emergency event. Moreover, integrating AHMS at fog provides continuous services to detect falls even during times when the Internet connection is interrupted. Furthermore, fog also sends the collected information to the AHMS at the cloud for storage and continuous data analysis purposes.

In [21], the Emergency Help Alert Mobile Cloud (E-HAMC) system was proposed to reduce the delay in sending an emergency alert to the hospital. Therefore, the method used fog to process the health data and a smartphone to send a GPS location of the event taking place. An experiment was conducted, with the results showing that the delay in sending the emergency alert with fog computing was lower compared to sending it directly from the cloud.

In [73], a smart F2C adaptor (SF2CA) was proposed to develop a monitoring system for COPD patients that provides real-time response to adjust the oxygen dose dynamically. Based on the dynamicity level of the data, the SF2CA decides the location for processing and decision making. The location can be either at the fog (i.e. high dynamicity) or the cloud (i.e. low dynamicity). In the above work, the fog is used to store the environmental data from sensors. Meanwhile, the cloud is employed to store the history data of the patient and to implement a predictive model known as the quasi-static data model.

In [74], an Obstructive Sleep Apnea (OSA) monitoring system was proposed, consisting of three layers, namely the IoT layer, fog computing layer and cloud layer. In the above work, fog computing was implemented at the Smart IoT Gateway. Fog was incorporated into a complex event processing to pre-process the health-related data so that immediate action can be taken in case of abnormal detection. The pre-processed data was made available at the cloud to improve the administration of the data. The performance of the proposed OSA was evaluated in terms of latency where the Smart IoT Gateway operating at the fog computing layer used different Low Power Wireless Networks which are Bluetooth, ZigBee and IPv6 over Low Power WPAN (6LowPAN) protocol, [75]–[77]. The results showed that processing health data at the fog layer and sending the feedback from fog had lower latency compared to processing at the cloud layer and sending the feedback from the cloud. Moreover, the authors reported that the complexity of protocol integration also increased the latency. Their experiment showed that the 6LowPAN protocol had the highest latency, followed by Bluetooth and Zigbee.

The authors in [78] claimed that a single edge layer is not enough to serve health applications. Therefore, to increase the scalability of fog computing, two types of edge layers were proposed, namely fog layer and intermediate fog layer. The proposed intermediate fog layer aims to reduce the load overhead at the fog layer. In the above work, the fog layer was used to process and analyse the health data. Meanwhile, the intermediate fog layer filters the processed data to eliminate redundant data before what is left is sent to the cloud for permanent storage and analysis of geo health data.

A resource management technique, namely the QoS-aware Resource Allocation Algorithm, was proposed in [40] to reduce the congestion at the cloud by offering an efficient utilisation of resources in fog computing. In this work, a fog server manager was proposed to monitor the availability of fog resources. Moreover, the processing tasks can be divided into several subtasks as per resource availability. The job will be sent to the cloud for processing only if the resources at the fog are not available. A simulation using iFogSim toolkits was performed, and the results showed that the delay in processing the demand decreased when placing the virtual machine at the fog servers compared to the processing at the cloud. However, the authors also claimed that placing the virtual machines at the optimum locations in the fog devices would further decrease the delay.

In [53], fog was formed into several clusters. The inter-model communication latency was given higher priority when creating the cluster. Each fog cluster was responsible for a particular healthcare solution while each healthcare solution could be run in multiple clusters. The proposed fog-based architecture was tested via simulation, and the results showed that the average network delay was lower compared to the cloud-based solution. However, the results also revealed that the integration of cloud-fog was essential during high demands for health services.

A patient monitoring system for Ambient Assisted Living using fog computing (FAAL) was proposed in [79] to observe neurological disorders in people with Ambient Assisted Living. The system utilised a k-means clustering algorithm to reduce the load on the communication infrastructure. In the above work, fog computing was utilised to perform data cleaning, segmenting, analysis and to send an alert signal during the emergency event. Meanwhile, the cloud was used to perform classification, prediction and education based on the data collected by the BANs transmitted via the fog gateway. The results showed that the FAAL system had a low response time compared to the conventional approach where the cloud is used to perform the analysis and to send the alert signal.

A hierarchical computing architecture, HiCH, was proposed in [80] to overcome the limited computation at the edge node. In the above work, the features offered by both fog and cloud computing were combined in HiCH. The proposed architecture used the MAPE-K model, which has four different components, namely monitor, analyse, plan and execute. Each component shared the system knowledge to manage the system resources efficiently. To ensure that each component was properly mapped into the HiCH architecture, system management was integrated into the system. The system management will periodically tune the computing components based on the input and the computations requirement of the model. The cloud performs the heavy computations, and the output is shared with the fog to perform the plan and the execution services. The fog node also reduces the traffic by filtering the redundant data from the sensor nodes. Moreover, the plan and execution can still be performed during times when Internet connection is interrupted. The proposed architecture was demonstrated by performing continuous health monitoring to assess cardiovascular patients, and the results showed that the proposed HiCH had low response time and high traffic reduction.

A medical warning system which uses the concept of fog computing was proposed in [10]. The work used autonomic computing proposed by IBM Corporation, which consists of four components, namely monitor, analyse, plan, and execute, to analyse the health data and decision making. Fog can be used to process the ECG data and to perform decision making even during times when Internet connection is interrupted. This is achieved by shifting the plan and executing components from the cloud to the local gateway (fog). An experiment using two types of fog devices with different processing capabilities, i.e. Raspberry Pi Zero and Jetson-TK1, was demonstrated with the proposed architecture to measure the latency due to transmitting data from the sensor node to the gateway, processing at the gateway and sending feedback from the gateway to the user. Note that, the Raspberry Pie Zero is a small board with 1 GHz processing capability while the Jetson-TK1 is a significantly more powerful board with a quad-core 2.32 GHz processing. The results were then compared with the traditional approach, where everything is processed at the cloud. The results indicated that the latency of using Jetson-TK1 as the fog gateway was lower than with the traditional approach. Meanwhile, the Raspberry Pi Zero had higher latency than the traditional approach, mainly due to the limited processing power, which increases the computation time.

The Health Level Seven (HL7) standard [81] has been widely studied to enable the interoperability of healthcare information between large health institutions. Due to this, a framework that standardises the exchange of health information between healthcare entities by using the HL7 standard was proposed in [82]. In this work, the data translation from plain and XML data formats to the HL7 standard was performed by introducing a new software tool. Meanwhile, Iguana/Chameleon tool [83], an application that is used to convert data into various formats besides allowing users to receive, transform and exchange any wanted data, was used to benchmark the implementation. The work evaluated the integration of the HL7 standard into the eHealth system for lightweight devices (sensor device) that used Zigbee and Wi-Fi communication protocol. The results revealed that converting data from the plain and XML formats to the HL7 increased the size of the data. This, in turn, increased the transmission time needed to send the data from the lightweight devices to the fog for further processing. Due to this, the above-mentioned work suggested that data translation of medical records exchange should be performed in the fog infrastructure rather than in a lightweight device. Besides, they also indicated that the Wi-Fi protocol is the best option to transmit a high number of medical records.

2.6.3 Accuracy

Accuracy is one of the most important parameters in health monitoring applications. Data accuracy allows doctors to give proper treatment to the patients. In [70], the accuracy of determining the state of event of the patient's condition was tested by integrating the classifier algorithms at the fog layer. Several classifier algorithms, including the Bayesian Belief Network (BBN), neural network, k-nearest neighbour and linear regression, were tested. The results revealed that the BBN had the highest accuracy compared to other algorithms. Meanwhile, in [84], the authors claimed that the J48 algorithm which generates a classification-decision tree for the given data set by recursive partitioning data [85,][86], had high accuracy. Therefore, the algorithm was used at the fog to classify the category of infection of the users who have been infected by the Chikungunya virus. The results indicated that the system is capable of detecting the risk-prone regions which have been infected with the virus.

In [71], a new fall detection algorithm based on acceleration magnitude values and non-linear time series analysis techniques was developed and implemented in the authors' proposed U-Fall system. The system was tested and the results showed that the developed algorithm had a lower missing rate and lower false alarm rate than the existing algorithm i.e. threshold-based technique, T-system and pattern matching system, and P-system.

A complex learning algorithm is used in the HiCH architecture to perform data analytics in the cloud [80]. In said work, the data set obtained from the cloud was shared with the fog to perform the plan and execution. The proposed architecture was demonstrated by performing continuous monitoring of cardiovascular patients, and the results showed that sharing the data set between the cloud and fog provided high accuracy.

A smart system to monitor patients with OSA was proposed in [74]. In the work, the fog layer performed the pre-processing of health data to detect any abnormalities in the health condition of patients so that the doctor can be notified in real time. Meanwhile, cloud computing was used to perform descriptive analysis, and this involved sending a batch of data processing to the cloud layer to ascertain the behaviour of said data and perform a predictive analysis for the development of services. This data included the pre-processed data at the fog layer and the open data catalogue which is available in smart cities. The aim of having the descriptive and predictive analysis is to help the doctor decide if the treatment given to the patient should be changed based on the health evolution of the patient for accuracy purposes. The proposed system also demonstrated that the prediction of the air quality index (AQI) provides 93.3% effectiveness.

2.6.4 Cost

Cost is one of the parameters which must be addressed in developing a health monitoring system. However, developing a smart health system at low cost is very challenging, especially when it comes to meeting the service demands from users while providing high-quality services. These challenges include operating expenses (OPEX) and capital expenditure (CAPEX).

The proposed privacy-preserving fog-assisted information sharing scheme (PFHD) in [60] achieved lightweight encryption on devices by offloading part of the encryption cost from the devices to the fog servers. The aim of this offloading was to lighten the burden on sensor devices in terms of computation and storage cost to perform efficient encryption for data privacy preservation. They compared the performance of PFHD at fog with cipher text-policy attribute-based encryption (CP-ABE) [87] and the results indicated that the PFHD was more efficient than CP-ABE in terms of computation and storage cost. Meanwhile, the type of security design for authentication purposes between the sensor and server also plays an important role in reducing the computational cost of the system. In [88], the hash function and secret key cryptosystem were designed to secure the communication between the sensor node and the fog server. The designed security protocol under SecHealth showed that the computation cost of the sensors in fog was low during the authentication phases compared to the existing protocol in [89]. The work in [90] also utilised fog to perform the authenticating and authorising protocol to reduce the burden on the sensor node and cloud.

A fog computing architecture, SmartFog was proposed in [91], which utilised low-resource machine learning on the fog node to analyse the pathological speech data from smart watches worn by patients with Parkinson’s Disease. In the above work, an acoustic analysis software program, together with Praat scripting language [92], was used to extract the features based on the pitch (frequency) and intensity of the collected samples, which comprised sound files with utterances. The analysis of the features was performed at the fog device by using the k-means clustering algorithm, which employs Python programming language. The proposed SmartFog architecture was tested using Intel Edison and Raspberry Pi, with the results showing that the Raspberry Pi outperformed the Intel Edison in terms of runtime and average CPU usage. A simulation using iFogSim tools to compare the total instances cost with the increasing number of applications request services to be performed in cloud and fog was performed in [53]. The results indicated that fog offered less service charge compared to the cloud. This is because fog uses micro computing instances (MCIs) which can be adjustable according to the number of application modules requested. Meanwhile, the cloud-based solution uses a virtual machine (VM), where the configuration is predefined. Due to this, the service charge for MCIs is based on the context of the module, while for VMs, the full-service charge is required, whatever the usage is.

In [65], the cost of deploying the fog nodes was reduced by using Field-Programmable Gate Array (FPGA) technology. The authors also claimed that this was the first work to employ the FPGAs technology in fog at the infrastructure and architecture level. In [72], the authors claimed that the medical and healthcare costs can be reduced if the life-threating events are detected earlier so that immediate action can be taken to save patients. Therefore, they proposed the Automatic Healthcare Management System

(AHMS), which continuously monitors and analyses the health data provided by the wearable devices and personal health records.

2.7 Conclusions

Recent times have seen the remote health monitoring system gain a great deal of attention due to increases in the ageing population and chronic diseases. There exists significant research on the fog-cloud based health monitoring system and IoT devices to support the development of healthcare applications with low energy consumption, low latency, high accuracy, high security and low cost. Although most of the parameters have been taken into consideration to develop the healthcare system, the energy which the network architecture consumes when deploying the health monitoring system has not been emphasised. Therefore, in the present study, the performance of fog computing for health monitoring applications is investigated in terms of the energy consumption of networking equipment and processing.

Chapter 3

A proposed health monitoring system with fog computing architecture

3.1 Introduction

There are several works that have considered the use of fog computing for health monitoring applications. However, the essential aspect of the energy consumption in transport networks and the impact of the fog locations in the edge network on energy efficiency has not been studied. This chapter proposes a new framework for an energy-efficient health monitoring system that performs real-time monitoring in a patient-centred environment by leveraging the concept of fog computing. We present the proposed health monitoring system with fog computing and explain the functions of the modules at the fog and cloud layer. Also, we introduce the proposed fog computing architecture for health monitoring applications using a Gigabit Passive Optical Network (GPON) access network, which is considered in this work due to its energy-efficiency. We explain in detail the purposes of each layer in the proposed fog architecture and introduce the candidate locations of the fog at the access network. A Mixed Integer Linear Programming (MILP) model is developed using AMPL software with CPLEX 12.8 solver as a platform to optimise the locations of the processing servers at the access layer so that the energy consumption of both networking equipment and processing are minimised. We also developed a MILP model for the conventional approach as a benchmark to evaluate the performance of the proposed fog approach in terms of the networking equipment and processing energy consumption. Note that, the MILP approach is chosen as the optimisation technique in this work rather than the other optimisation models available as in listed in [93] because the linear constraints related to the linear programming subproblem result in a convex feasible region, which is guaranteed to obtain the global optimum [94].

3.2 Health monitoring system with fog computing

This section presents the proposed architecture of the fog-based health monitoring system, which is divided into three modules: i) health data analysis and decision-making module, ii) fog storage module, and iii) cloud storage module, as illustrated in Figure 3.1. The health data analysis and decision-making module and the fog storage module are embedded in the fog layer, whereas the cloud storage module is incorporated in the cloud layer.

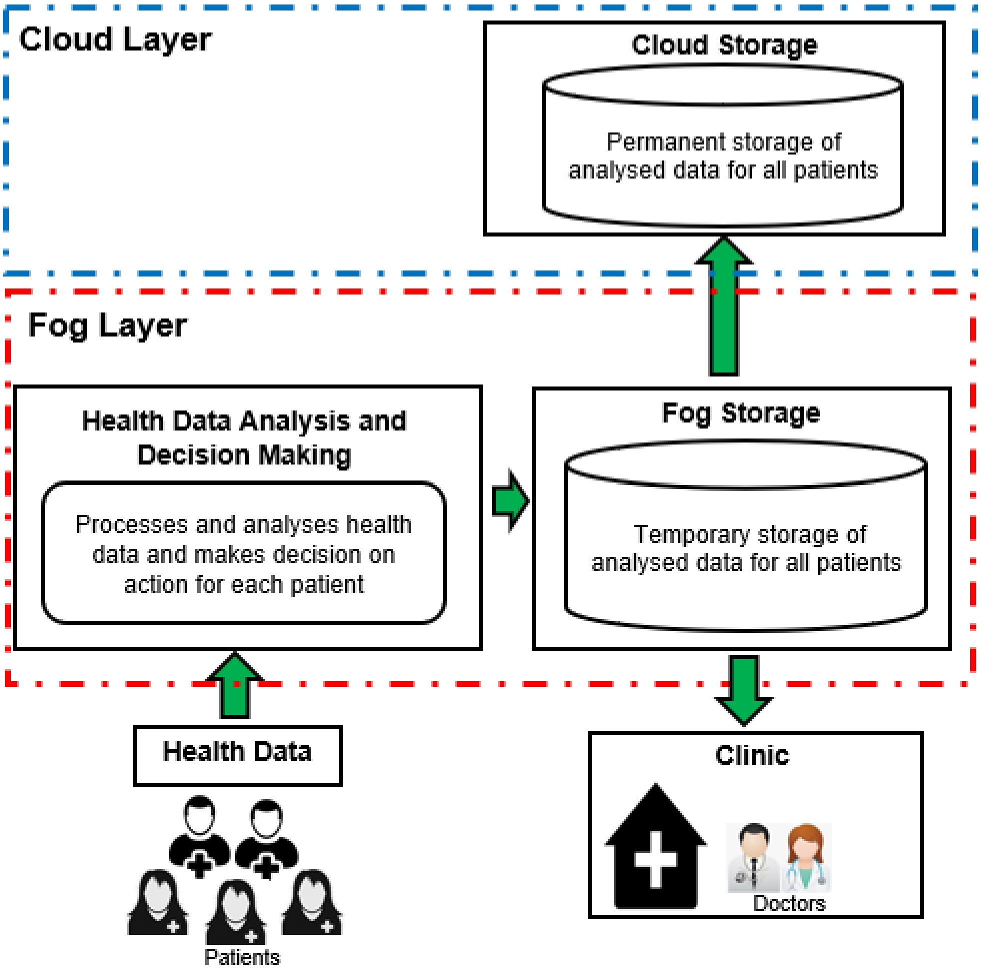

Cloud Layer
Cloud Storage
Permanent storage of analysed data for all patients
Fog Layer
Health Data Analysis and Decision Making
Processes and analyses health data and makes decision on action for each patient
Fog Storage
Temporary storage of analysed data for all patients
Health Data
Patients
Clinic
Doctors

Figure 3.1: Architecture of the proposed system

Below we explain in detail the function of each module:

1. The health data analysis and decision-making module performs three tasks. The first is aggregating health data derived from multiple patients via wireless-connected devices, for example, smartphones to monitor postoperative atrial fibrillation (AF). The second task processes and analyses the health data of each patient and matches it with the disease symptom based on the extracted features of the health data. The final task is making decisions on the action taken against irregular physical data of

the patients, such as informing the emergency medical service resources to act fast on patients who seek aid. Nonetheless, in some cases, the doctors would re-diagnosis the results before making the final decision.

2. The fog storage module serves as a temporary storage for health results besides providing accessibility for patient and doctors upon pressing demands. This module is also used to send the analysed health data to the cloud storage and the clinic for permanent storage and feedback purposes, respectively.
3. The cloud storage module permanently stores the analysed results of patients for history purposes. This module offers accessibility for both patients and doctors, similar to that in the fog storage module.

3.3 Fog computing architecture for health monitoring applications with Gigabit passive optical network (GPON) access network

The architecture of using the GPON network is characterised by four networking layers, as portrayed in Figure 3.2. It is worth noting that a redundant connection is present for each device to increase the resilience of the network.

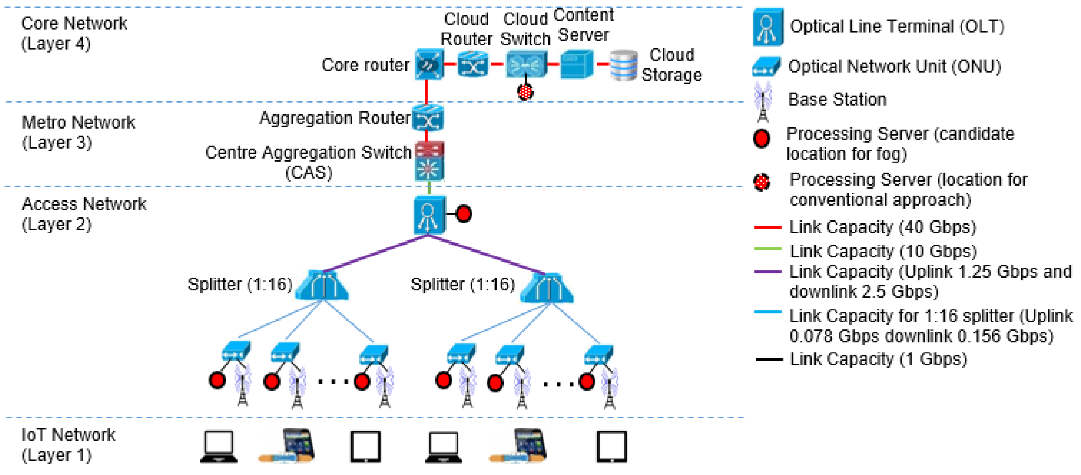

Core Network
(Layer 4)
Cloud Router
Cloud Switch
Content Server
Core router
Cloud Storage
Metro Network
(Layer 3)
Aggregation Router
Centre Aggregation Switch (CAS)
Access Network
(Layer 2)
Splitter (1:16)
Splitter (1:16)
IoT Network
(Layer 1)
Optical Line Terminal (OLT)
Optical Network Unit (ONU)
Base Station
Processing Server (candidate location for fog)
Processing Server (location for conventional approach)
Link Capacity (40 Gbps)
Link Capacity (10 Gbps)
Link Capacity (Uplink 1.25 Gbps and downlink 2.5 Gbps)
Link Capacity for 1:16 splitter (Uplink 0.078 Gbps downlink 0.156 Gbps)
Link Capacity (1 Gbps)

Figure 3.2: GPON architecture in the fog network

The first layer (i.e. Layer 1) is the bottom-most layer that is comprised of IoT devices, mobile phones, iPads, etc. which supports Machine-to-Machine (M2M) communication devices with a connection to wireless body sensors to both monitor the health of patients and to send data to the network. The second layer (i.e. Layer 2) is the access layer that serves as the fog computing layer. This layer aggregates data from layer 1 via gateways such as a LTE-M base station (1.4MHz bandwidth for LTE-M), Wi-Fi access point, etc. It also consists of fog servers that can process, analyse, and perform temporary storage, Optical Network Units (ONUs), and an Optical Line Terminal (OLT). Fog computing processing resources serving the health monitoring application can be deployed at ONUs and the OLT. Placing the processing servers (PSs) at ONUs, which is closer to the users, decreases the energy consumption of networking equipment, however, it will increase the required number of PSs. Meanwhile, utilising PSs at OLT reduces the number of PSs required as it is a shared point between the access points although will increase the energy consumption of the networking equipment. In fact, the fog server performs the processing, web server, firewalls, and security functions. Next, the gateway and the fog are connected via the ONU, which has an internal access switch. The ONUs are connected to the OLT via a passive splitter that converts the electrical signal to an optical signal before forwarding them to the OLT. The connections between the gateway, fog, and ONU are of copper wire, and an optical fibre between the ONU and OLT.

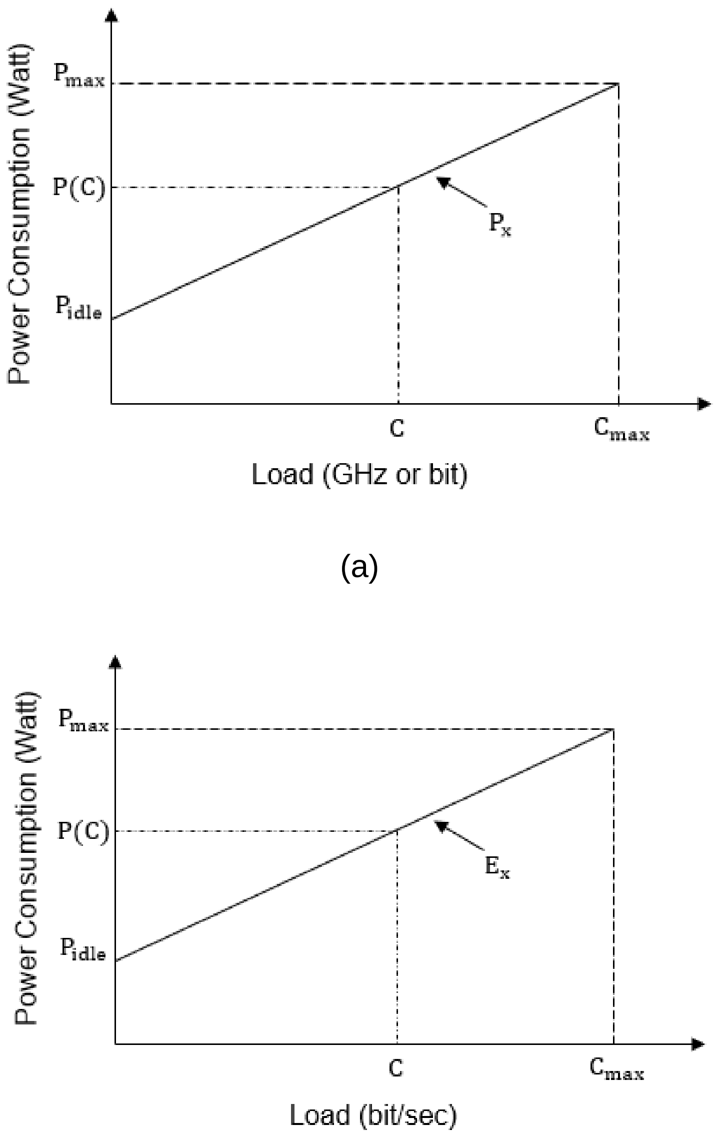
The third layer (i.e. Layer 3) is the metro layer that consists of a centre aggregation switch (CAS) and aggregation router. The CAS aggregates and fast-forwards data between the fogs in the access network. The aggregation router serves as a gateway to connect the access network to the core network. Connections between the devices in this layer are made of optical fibre. The fourth layer (i.e. Layer 4) is the upper-most layer in the architecture that has IP over WDM core network devices, such as core routers. This layer is integrated with the central cloud that consists of cloud routers, cloud switches, content servers, and cloud storage. The central cloud is used to permanently store the health data of patients, which is significant in health application [95]. Note that, for the conventional approach, the processing server is located at the cloud switches.

3.3.1 Link capacity considerations in the network for health monitoring applications

The total IP traffic of M2M communication from the global IP traffic in 2016 was 2%, and is expected to be 5% in 2021 [96]. Cisco also reported that the total M2M connected devices and the connected health consumers of M2M connection in 2015 were 4.9 billion and 144 million, respectively, and are estimated to hike up to 12.2 billion and 729 million, respectively, in 2020 [97]. This signifies that the M2M connected devices for healthcare applications were approximately responsible for 3% of the traffic in 2015 and will be responsible for 6% in 2020. For this work, only 5% had been employed to predict M2M traffic from the global traffic, while 6% represented healthcare traffic from the total M2M traffic. These calculated percentages are used to estimate the link capacities in the network for healthcare applications. Furthermore, it is worth noting that the link capacities at all layers (access, metro and core layers) are considered to serve traffic for all other applications and not only M2M applications. Thus, the link capacities dedicated to healthcare application at all layers are 0.3% of the maximum capacities. This can be explained by noting that 5% of all IP traffic is M2M traffic, while 6% of the M2M allocation is for the healthcare application. Note, all 0.3% of the maximum link capacities are considered to be available for our healthcare application.

3.3.2 Power profile of networking and processing equipment for health monitoring applications

The power consumption for most networking and computing devices reflects a linear power profile [98]. Hence, power consumption of all networking equipment and PS consists of both an idle and a linear proportional part.



(a)

(b)

Figure 3.3: Power consumption model for (a) processing servers and cloud storage (b) other networking devices

Figure 3.3-(a) illustrates the power profile for the PS and cloud storage while Figure 3.3-(b) illustrates the power profile for the other networking equipment. The aspect of power consumption had been determined based on the fixed idle power and the load dependent power. Equation (3-1) calculates the linear form power consumption of the PS and the cloud storage, where P_{idle} denotes its idle power, while the graph slope, (P_x) refers to power per

GHz for PS and power per Gbit for cloud storage. C refers to the offered load found in GHz and Gbit for PS and cloud storage, respectively.

$$P(C) = P_{idle} + C\,\frac{P_{max} - P_{idle}}{C_{max}} = P_{idle} + C\,P_x \qquad (3\text{-}1)$$

Meanwhile, Equation (3-2) calculates the linear form of other networking devices power consumption, where P_{idle} denotes the idle power while the slope of the graph (E_x) reflects the increased energy per bit. Besides, C denotes the offered load in bit per second.

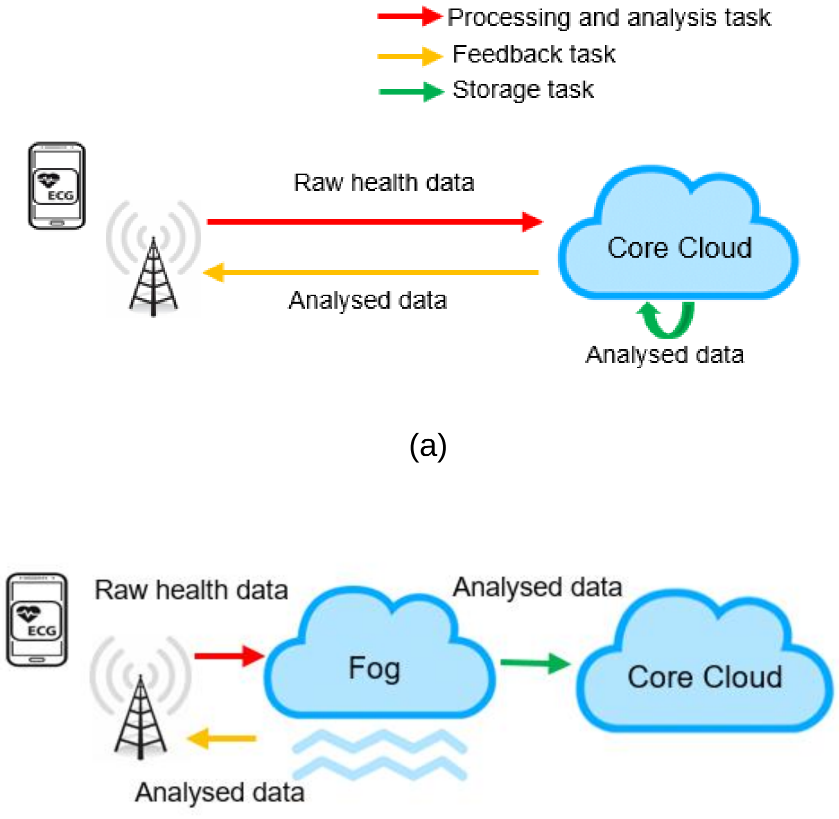
$$P(C) = P_{idle} + C\,\frac{P_{max} - P_{idle}}{C_{max}} = P_{idle} + C\,E_x \qquad (3\text{-}2)$$

The maximum power consumption of the networking equipment and the PS, together with their maximum capacity used to calculate both idle and load dependent power, can be retrieved from data sheets and references. As for ONU, the maximum capacity, $CONU$, was considered as the summation of the maximum uplink capacity, 1.25 Gbps [99], and maximum downlink capacity, 2.5 Gbps [99], to obtain E_b. Note that, the networking devices are shared by multiple applications while the considered PSs are dedicated to the healthcare application. As previously discussed, the healthcare application is thus considered to contribute to 0.3% of the idle power of the networking devices. Also, we considered our healthcare application to be responsible for 0.3% of the idle power of the networking equipment. Moreover, due to cooling and other overheads in the network devices, such as uninterruptable power system at network sites, a power usage effectiveness (PUE) factor is incorporated. PUE is defined as the ratio of the total power used for equipment, cooling and other overheads to the power used by the equipment (communication or computing equipment). In network infrastructure, a typical telecom office PUE is 1.5 [100], [101]. Therefore, for IP over WDM, metro, and access networks, a PUE of 1.5 is considered [102], [103]. Meanwhile, based on the United State data center energy usage, data centers PUE varies based on the size of data centers as more efficient cooling equipment are used in larger data centers [104]. The typical data centers PUE varies between 1.1 for the large data center to 3 for small data centers [104], [101]. In this work, we adopted a PUE of 2.5 for small distributed clouds [105]. In addition, a PUE of 2.5 was set for the fog.

3.4 System flow of health monitoring applications in the network

In this work, the health data of each patient was monitored. Two approaches were incorporated in this study, which are, the conventional approach (CA) and the proposed approach with fog optimisation (FOA). There are three tasks to be carried out in each approach; processing and analysis, feedback and storage. The processing task extracts health data features, for instance, ECG signal heart rate and QRS duration (Q, R and S label the start, peak and end of a heart beat pulse), which are essential in detecting AF among patients, are determined through the processing task. This is performed by the PS which is located at the central cloud in CA, while at fog in FOA.

Next, the analysis process refers to the diagnosis of health data features (i.e. heart rate and QRS duration for ECG signal) so as to monitor the health condition among patients, either normal or abnormal, to decide the appropriate actions given to the patients. The analysis process is performed at the same PS that performs the processing. Then, the feedback task is performed by sending the analysed data to the clinics. The last process performs permanent storage of the analysed data of the patients at the cloud. However, in FOA, the fog is used first to temporarily store the analysed data of the patients before being stored permanently at cloud storage.



(a)

(b)

Figure 3.4: System flow of (a) conventional approach (CA) (b) proposed approach (FOA)

Figure 3.4-(a) and Figure 3.4-(b) illustrate the system flow of the CA and the proposed approaches (FOA), respectively. In CA, the raw health data are sent to the central cloud for processing and analysis, feedback and permanent storage. It is important to note that the PS at the central cloud also stores the latest analysed health data of each patient for the next analysis. Meanwhile, in FOA, both processing and analysis of raw health data and the feedback task are performed at the fog. In fact, fog servers are also used to temporarily store the analysed health data, which are later sent to the central cloud for permanent storage.

3.5 Mathematical model for Energy-efficient fog computing health monitoring applications with LTE-M (EEFC)

This section presents the Mixed Integer Linear Programming (MILP) model that has been developed for fog approach (FOA) to minimise the energy consumption in both networking and processing equipment by optimising the location of PS at access network. Note, the energy consumption of networking equipment includes the energy consumed by all networking devices at all layers while the processing energy consumption refers to the energy consumed by the processing servers. Before introducing the model, we define the sets, parameters and variables used as in Table 3.1 (also can be found in Appendix 1). Note that, the nodes refer to the networking devices at all layers while the candidate nodes refer specifically to the nodes that can be used to host the PS. The mathematical source code of the EEFC model can be found in Appendix 2.

Table 3.1: The sets, parameters and variables used in MILP

Sets	
CL	Set of clinics
BS	Set of BSs
ONU	Set of ONUs
OLT	Set of OLTs
CAS	Set of centre aggregation switches
AR	Set of aggregation routers
CR	Set of core routers
CLR	Set of cloud routers
CLS	Set of cloud switches
CS	Set of content servers
CST	Cloud storage
N_m	Set of neighbouring nodes of node m in the network
N	Set of nodes ($N \in CL \cup BS \cup ONU \cup OLT \cup CAS \cup AR \cup CR \cup CLR \cup CLS \cup CS \cup CST$)
FN	Set of candidate locations to deploy PS (fog) ($FN \in ONU \cup OLT$)
Parameters	
$s\ and\ d$	Denote source node s and destination node d of traffic between a node pair
$i\ and\ j$	Denote end nodes of a physical link in the network, $i, j \in N$
Pt_s	Number of patients in clinic s
IBS	Idle power consumption of a base station (W)
PBS	Power per physical resource block (PRB) of a base station (W/PRB)

R	Maximum number of PRBs in a base station dedicated for healthcare applications
$PONU$	Maximum power consumption of an ONU (W)
$IONU$	Idle power consumption of an ONU (W)
$CONU$	Maximum capacity of an ONU (bps)
$POLT$	Maximum power consumption of an OLT (W)
$IOLT$	Idle power consumption of an OLT (W)
$COLT$	Maximum capacity of an OLT (bps)
$PCAS$	Maximum power consumption of a centre aggregation switch (W)
$ICAS$	Idle power consumption of a centre aggregation switch (W)
$CCAS$	Maximum capacity of a centre aggregation switch (bps)
PAR	Maximum power consumption of an aggregation router (W)
IAR	Idle power consumption of an aggregation router (W)
CAR	Maximum capacity of an aggregation router (bps)
PCR	Maximum power consumption of a core router (W)
ICR	Idle power consumption of a core router (W)
CCR	Maximum capacity of a core router (W)
$PCLR$	Maximum power consumption of a cloud router (W)
$ICLR$	Idle power consumption of a cloud router (W)
$CCLR$	Maximum capacity of a cloud router (bps)
$PCLS$	Maximum power consumption of a cloud switch (W)
$ICLS$	Idle power consumption of a cloud switch (W)
$CCLS$	Maximum capacity of a cloud switch (bps)
PCS	Maximum power consumption of a content server (W)
ICS	Idle power consumption of a content server (W)
CCS	Maximum capacity of a content server (bps)

$PCST$	Maximum power consumption of a cloud storage (W)
$ICST$	Idle power consumption of a cloud storage (W)
$CCST$	Maximum capacity of a cloud storage (bit)
PPS	Maximum power consumption of a processing server (W)
IPS	Idle power consumption of a processing server (W)
Ωmax	Maximum number of patients per processing server
Λmax	Maximum storage capacity of processing server (bit)
δa	Data rate per patient to send raw health data from clinic to processing server (bps)
τa	Transmission time per patient to send raw health data from clinic to processing server (s)
Ra	Physical resource block per patient to send raw health data from clinic to processing server
α	Size of analysed health data per patient (bit)
δb	Data rate per patient to send analysed health data from processing server to clinic (bps)
τb	Transmission time per patient to send analysed health data from processing server to clinic (s)
Rb	Physical resource block per patient to send analysed health data from processing server to clinic
δc	Data rate per patient to send analysed health data from processing server to cloud storage (bps)
τc	Transmission time per patient to send analysed health data from processing server to cloud storage (s)
δ_{sd}	$\delta_{sd} = 1$ to send the storage traffic from processing servers located at candidate node s, to the cloud storage node d, $s \in FN, d \in CST$

x	Fraction of idle power consumption of networking equipment contributed by the healthcare application under consideration
λ_{ij}	The capacity of link ij dedicated for the healthcare application under consideration (bps)
η	Power usage effectiveness (PUE) of the access network, metro network and IP over WDM network
c	Power usage effectiveness (PUE) of the fog (processing server) and cloud equipment
M	A large enough number
Variables	
P_{sd}	Raw health data traffic from source node s to destination node d (bps), $s \in CL, d \in FN$
P_{ij}^{sd}	Raw health data traffic from source node s to destination node d that traverses the link between nodes i and j (bps), $s \in CL, d \in FN,\ i,j \in N$
P_i	Total raw health data traffic that traverses node i (bps), $i \in N$
F_{sd}	Analysed health data feedback traffic from source node s to destination node d (bps), $s \in FN, d \in CL$
F_{ij}^{sd}	Analysed health data feedback traffic from source node s to destination node d that traverses the link between nodes i and j (bps), $s \in FN, d \in CL, i,j \in N$
F_i	Total analysed health data feedback traffic that traverses node i (bps), $i \in N$
S_{sd}	Analysed health data storage traffic from source node s to destination node d (bps), $s \in FN, d \in CST$
S_{ij}^{sd}	Analysed health data storage traffic from source node s to destination node d that traverses the link between nodes i and j (bps), $s \in FN, d \in CST, i,j \in N$

S_i	Total analysed health data storage traffic that traverses node i (bps), $i \in N$
ω_{sd}	Number of patients from clinic s served by processing server located at candidate node d
Pa_{ij}	Number of patients in clinic i served by BS j to send raw health data traffic (integer)
Pb_{ij}	Number of patients in clinic i served by BS j to receive analysed health data feedback traffic (integer)
βa_j	Number of PRBs used in BS j to serve raw health data traffic (integer)
βb_i	Number of PRBs used in BS i to serve analysed health data feedback traffic (integer)
Y_d	$Y_d = 1,$ if a processing server is placed at candidate node $d,$ otherwise $Y_d = 0$, $d \in FN$
ϕ_d	Number of processing servers placed at candidate node d, $d \in FN$
τp_d	Processing and analysis time of processing server (seconds) at candidate node d, $d \in FN$
ζa_j	$\zeta a_j = 1,$ if raw health data traffic traverses node j, otherwise $\zeta a_j = 0$, $j \in N$
ζb_i	$\zeta b_i = 1,$ if analysed health data feedback traffic traverses node i, otherwise $\zeta b_i = 0$, $i \in N$
θc_i	$\theta c_i = 1,$ if analysed health data storage traffic traverses node i where node i is the source of a link, otherwise $\theta c_i = 0$, $i \in N$
ϑc_j	$\vartheta c_j = 1,$ if analysed health data storage traffic traverses node j where j is the end of a link, otherwise $\vartheta c_j = 0$, $j \in N$

ζc_i	$\zeta c_i = 1$, if the analysed health data storage traffic traverses node i where $\zeta c_i = \theta c_i\ OR\ \vartheta c_i$, otherwise $\sigma_i = 0$, $i \in N$
ν_i	ν_i is a dummy variable that takes value of $\theta c_i \oplus \vartheta c_i$, where $\oplus$ is an XOR operation, $i \in N$
EAN	Energy consumption of access network
$ETBS$	Total energy consumption of base stations
$EBSP$	Energy consumption of base stations required to relay raw health data traffic
$EBSF$	Energy consumption of base stations required to relay analysed health data feedback traffic
$ETONU$	Total energy consumption of ONUs
$EONUP$	Energy consumption of ONUs required to relay raw health data traffic
$EONUF$	Energy consumption of ONUs required to relay analysed health data feedback traffic
$EONUS$	Energy consumption of ONUs required to relay analysed health data storage traffic
$ETOLT$	Total energy consumption of OLTs
$EOLTP$	Energy consumption of OLTs required to relay raw health data traffic
$EOLTF$	Energy consumption of OLTs required to relay analysed health data feedback traffic
$EOLTS$	Energy consumption of OLTs required to relay analysed health data storage traffic
EMN	Energy consumption of metro network
$ECASS$	Energy consumption of centre aggregation switches required to relay analysed health data storage traffic

$EARS$	Energy consumption of aggregation routers required to relay analysed health data storage traffic
ECN	Energy consumption of core network
$ECRS$	Energy consumption of core routers required to relay analysed health data storage traffic
ECL	Energy consumption of cloud
$ECLRS$	Energy consumption of cloud routers required to relay analysed health data storage traffic
$ECLSS$	Energy consumption of cloud switches required to relay analysed health data storage traffic
$ECSS$	Energy consumption of content servers required to relay analysed health data storage traffic
$ECSTS$	Energy consumption of cloud storage required to store the analysed health data storage traffic
EFN	Energy consumption of fog nodes
EPS	Energy consumption of processing servers

We start by defining the energy consumption of the network (i.e. access, metro and core) and processing servers:

a) Energy consumption of access network, EAN:

The energy consumption of access network, EAN, is composed of the LTE base stations', ONUs' and OLTs' energy consumption. The energy consumption of access network is given in Equation (3-3):

$$EAN = (ETBS + ETONU + ETOLT)\ \eta \tag{3-3}$$

where η is the network PUE. The energy consumption of access network is composed of the energy consumed by three different tasks i.e. processing, feedback and storage tasks. In the processing task, raw health data is sent from the clinic to the PS at fog. In the feedback task, analysed health data is sent from the PS to the clinics. Meanwhile, in the storage task, analysed health data is sent from the PS to the cloud storage. Note that the three tasks occur at different times. The energy consumption of BS $(ETBS)$ is given as:

$$ETBS = EBSP + EBSF \tag{3-4}$$

where

$$EBSP = \sum_{i \in BS} (IBS\ x\ \zeta a_i\ + PBS\ \beta a_i)\ \tau a \tag{3-5}$$

$$EBSF = \sum_{i \in BS} (IBS\ x\ \zeta b_i + PBS\ \beta b_i)\ \tau b \tag{3-6}$$

The energy consumed by LTE BS, $ETBS$, is based on the number of PRBs (βa_i and βb_i) and the time the BS (τa and τb) is used to relay raw health data traffic, $EBSP$ and analysed health data feedback traffic, $EBSF$, as depicted in Equation (3-5) and Equation (3-6), respectively.

The energy consumption of the ONUs is given as:

$$ETONU = EONUP + EONUF + EONUS \tag{3-7}$$

where

$$EONUP = \sum_{i \in ONU} \left(IONU\ x\ \zeta a_i + P_i\ \frac{(PONU - IONU)}{CONU} \right)\ \tau a \tag{3-8}$$

$$EONUF = \sum_{i \in ONU} \left(IONU\ x\ \zeta b_i + F_i\ \frac{(PONU - IONU)}{CONU} \right)\ \tau b \tag{3-9}$$

$$EONUS = \sum_{i \in ONU} \left(IONU\ x\ \zeta c_i + S_i\ \frac{(PONU - IONU)}{CONU} \right)\ \tau c \tag{3-10}$$

The energy consumption of ONUs is calculated based on relaying the raw health data traffic, analysed health data feedback traffic and analysed health data storage traffic, as presented in Equation (3-7). The energy consumed by ONUs is proportional to the size of traffic traversing them and the utilisation time. Equations (3-8)-(3-10) depict the calculation of energy consumed by the ONUs to relay raw health data traffic, $EONUP$, analysed health data feedback traffic, $EONUF$ and analysed health data storage traffic, $EONUS$, respectively.

The energy consumption of the OLTs is given as:

$$ETOLT = EOLTP + EOLTF + EOLTS \tag{3-11}$$

where

$$EOLTP = \sum_{i \in OLT} \left(IOLT\ x\ \zeta a_i + P_i\ \frac{(POLT - IOLT)}{COLT} \right)\ \tau a \tag{3-12}$$

$$EOLTF = \sum_{i \in OLT} \left(IOLT\ x\ \zeta b_i + F_i\ \frac{(POLT - IOLT)}{COLT} \right) \tau b \tag{3-13}$$

$$EOLTS = \sum_{i \in OLT} \left(IOLT\ x\ \zeta c_i + S_i\ \frac{(POLT - IOLT)}{COLT} \right) \tau c \tag{3-14}$$

The energy consumed by the OLTs is based on relaying the three types of traffic explained for ONUs as depicted in Equation (3-11). The energy consumed by OLT to relay the traffic is proportional to the size of traffic traversing them and the utilisation time. Equations (3-12)-(3-14) depict the energy consumed by the OLT to relay raw health data traffic, $EOLTP$, analysed health data feedback traffic, $EOLTF$ and analysed health data storage traffic, $EOLTS$, respectively.

b) Energy consumption of metro network, EMN

The energy consumption of metro network, EMN, is composed of the energy consumption of centre aggregation switches and aggregation routers. Note that these devices are only used to relay the analysed health data storage traffic as the candidate locations of PS is at the access layer. Hence the raw health data traffic and analysed health data feedback traffic does not traverse the metro network. The energy consumption of the metro network is as given in Equation (3-15):

$$EMN = (ECASS + EARS)\ \eta \tag{3-15}$$

where

$$ECASS = \sum_{i \in CAS} \left(ICAS\ x\ \zeta c_i + S_i\ \frac{(PCAS - ICAS)}{CCAS} \right) \tau c \tag{3-16}$$

$$EARS = \sum_{i \in AR} \left(IAR\ x\ \zeta c_i + S_i\ \frac{(PAR - IAR)}{CAR} \right) \tau c \tag{3-17}$$

The energy consumed by the centre aggregation switches and aggregation routers are proportional to the size of traffic traversing them and the utilisation time as shown in Equation (3-16) and Equation (3-17), respectively.

c) Energy consumption of core network, ECN

The energy consumption of core network, ECN, is composed of the energy consumption of core routers as given in Equation (3-18):

$$ECN = ECRS\ \eta \tag{3-18}$$

where

$$ECRS = \sum_{i \in CR} \left(ICR\ x\ \zeta c_i + S_i\ \frac{(PCR - ICR)}{CCR} \right) \tau c \tag{3-19}$$

The energy consumption of core routers is based on the size of traffic traversing them and the utilisation time to relay the analysed health data storage traffic as shown in Equation (3-19).

d) Energy consumption of cloud, ECL

The energy consumption of cloud, ECL, is composed of energy of cloud routers, cloud switches, content servers and cloud storage. Note that the cloud storage is used to perform the storage task while other devices are only used to relay the analysed health data storage traffic. The energy consumption of the cloud is given in Equation (3-20):

$$ECL = (ECLRS + ECLSS + ECSS + ECSTS)\ c \qquad (3\text{-}20)$$

where c is the cloud PUE. The energy consumption of cloud routers, $ECLRS$, cloud switches, $ECLSS$, content severs, $ECSS$ and cloud storage, $ECSTS,$ are given as:

$$ECLRS = \sum_{i \in CLR} \left(ICLR\ x\ \zeta c_i + S_i\ \frac{(PCLR - ICLR)}{CCLR} \right) \tau c \qquad (3\text{-}21)$$

$$ECLSS = 2 \sum_{i \in CLS} \left(ICLS\ x\ \zeta c_i + S_i\ \frac{(PCLS - ICLS)}{CCLS} \right) \tau c \qquad (3\text{-}22)$$

$$ECSS = \sum_{i \in CS} \left(ICS\ x\ \zeta c_i + S_i\ \frac{(PCS - ICS)}{CCS} \right) \tau c \qquad (3\text{-}23)$$

$$ECSTS = 2 \sum_{i \in CST} \left(ICST\ x\ \zeta c_i + S_i\ \tau c\ \frac{(PCST - ICST)}{CCST} \right) \tau c \qquad (3\text{-}24)$$

The energy consumption of cloud storage is calculated based on the size of analysed data stored in the cloud storage and the time the device is utilised, while the energy consumption of the other devices is based on the size of traffic traversing those devices and the time the devices are utilised to relay the analysed health data storage traffic. Note that the energy consumption of the cloud switches and the cloud storage are multiplied by ‘2’ for redundancy purposes [105].

e) Energy consumption of fog nodes, EFN:

The energy consumed by the fog, EFN, reflects the energy consumed by processing server, $EPS,$ as given below:

$$EFN = EPS\ c \tag{3-25}$$

where

$$EPS = \sum_{d \in FN} (IPS\ \phi_d\ (\tau a + \tau b + \tau c) + PPS\ \tau p_d) \tag{3-26}$$

The energy consumed by the processing servers is determined by considering the idle energy consumption and the energy consumed to perform the processing. The idle energy consumption is calculated by considering the following: the time to receive raw health data from clinic, τa, the time to transmit the analysed health data to clinics, $\tau b,$ as well as the time to transmit the analysed health data to cloud storage, τc. Note that the processing server always works at full utilisation, hence maximum power is consumed to process the raw health data. The energy consumption of processing and analysis for the processing server is determined by considering the time to perform the processing and analysis, τp_d. Note that the same processing servers are utilised in both fog and cloud.

The model is defined as follows:

Objective:

Minimise the total energy consumption of access network, EAN, metro network, EMN, core network, ECN, cloud network, $ECL,$ and processing server, EFN, given as:

$$EAN + EMN + ECN + ECL + EFN \tag{3-27}$$

Subject to:

1) Association of patients to a processing server.

$$\omega_{sd} \leq Pt_s\, Y_d \quad ; \ \forall s \in CL, \forall d \in FN \tag{3-28}$$

Constraint (3-28) is used to allocate a patient from clinic s, to a servers, namely to be served by the processing server located at node d. Note that, if a patient is allocated to a candidate location, this location should have a fog node in the fog approach.

$$\sum_{d \in FN} \omega_{sd} = Pt_s \quad ; \ \forall s \in CL \tag{3-29}$$

Constraint (3-29) ensures that all patients at clinic s are assigned to a processing server located at any node d.

2) Traffic from clinics to processing server.

$$P_{sd} = \omega_{sd}\, \delta a \quad ; \ s \in CL, d \in FN \tag{3-30}$$

Constraint (3-30) calculates the raw health data traffic from clinic s to the processing server located at node d based on the association of patients from clinic to processing server, ω_{sd}, as well as the data rate provisioned for each patient, δa, to perform the transmission.

3) Traffic from processing server to clinics.

$$F_{sd} = \omega_{ds}\, \delta b \quad ; \ \forall s \in FN, d \in CL \tag{3-31}$$

Constraint (3-31) calculates the analysed health data feedback traffic from the processing server located at node s, to clinic d. In fact, this is based on the total number of patients in the clinic served by the processing server at fog, ω_{ds} and the data rate provisioned for each patient, δb, to perform the transmission.

4) Traffic from processing server to cloud storage.

$$S_{sd} = \sum_{i \in CL} \omega_{is}\, \delta c\, \delta_{sd} \quad ; \ \forall s \in FN, d \in CST \tag{3-32}$$

Constraint (3-32) calculates the analysed health data storage traffic from processing server located at node s, to cloud storage d. Note that in this work we only utilise one cloud storage, hence, δ_{sd}=1. In fact, this is based on the total number of patients from clinic i served by the processing server at fog, ω_{is}, and the data rate provisioned for each patient, δc, to perform the transmission.

5) Flow conservation in the network.

$$\sum_{j \in Nm[i]: i \neq j} P_{ij}^{sd} - \sum_{j \in Nm[i]: i \neq j} P_{ji}^{sd} = \begin{cases} P_{sd}\ if\ i = s \\ -P_{sd}\ if\ i = d \\ 0\ otherwise \end{cases} \tag{3-33}$$

$$s \in CL, d \in FN, i \in N$$

$$\sum_{j \in Nm[i]: i \neq j} F_{ij}^{sd} - \sum_{j \in Nm[i]: i \neq j} F_{ji}^{sd} = \begin{cases} F_{sd}\ if\ i = s \\ -F_{sd}\ if\ i = d \\ 0\ otherwise \end{cases} \tag{3-34}$$

$$s \in FN, d \in CL, i \in N$$

$$\sum_{j \in Nm[i]:i \neq j} S_{ij}^{sd} - \sum_{j \in Nm[i]:i \neq j} S_{ji}^{sd} = \begin{cases} S_{sd} \; if \; i = s \\ -S_{sd} \; if \; i = d \\ 0 \; otherwise \end{cases} \tag{3-35}$$

$$s \in FN, d \in CST, i \in N$$

Constraints (3-33)-(3-35) ensure that the total incoming traffic is equivalent to the total outgoing traffic for all nodes in the network, except for the source and destination nodes for processing, feedback, and storage for tasks, respectively.

6) Total traffic traversing node.

$$P_i = \left(\sum_{s \in CL} \sum_{d \in FN:s \neq d} \sum_{j \in Nm[i]:\mathrm{i} \neq \mathrm{j}} P_{ji}^{sd} \right) \quad ; \;\; \forall i \in N \tag{3-36}$$

$$F_i = \left(\sum_{s \in FN} \sum_{d \in CL:s \neq d} \sum_{j \in Nm[i]:\mathrm{i} \neq \mathrm{j}} F_{ij}^{sd} \right) \quad ; \;\; \forall i \in N \tag{3-37}$$

$$S_i = \left(\sum_{s \in FN} \sum_{d \in CST:s \neq d} \sum_{j \in Nm[i]:\mathrm{i} \neq \mathrm{j}} S_{ji}^{sd} + \sum_{d \in CST:i \neq d} S_{id} \right) \quad ; \;\; \forall i \in N \tag{3-38}$$

Equations (3-36)-(3-38) calculate the total raw health data traffic, analysed health data feedback traffic, and analysed health data storage traffic that traverse node i, respectively.

7) Link capacity constraint.

$$\sum_{s \in CL} \sum_{d \in FN} P_{ij}^{sd} \leq \lambda_{ij} \quad ; \;\; \forall i \in N, \forall j \in Nm[i]: i \neq j \tag{3-39}$$

$$\sum_{s \in FN} \sum_{d \in CL} F_{ij}^{sd} \leq \lambda_{ij} \quad ; \;\; \forall i \in N, \forall j \in Nm[i]: i \neq j \tag{3-40}$$

$$\sum_{s \in FN} \sum_{d \in CST} S_{ij}^{sd} \leq \lambda_{ij} \quad ; \;\; \forall i \in N, \forall j \in Nm[i]: i \neq j \tag{3-41}$$

Constraints (3-39)-(3-41) ensure that the capacity of physical links used to send the total raw health data from all clinics s to processing servers at node d for processing task, the total analysed health data from all processing servers at node s to the clinic d for feedback task, and the total analysed health data from all processing servers at node s to the cloud storage d for storage task, respectively, does not exceed the maximum capacity of the links. Note that, as mentioned above, the three tasks occur at different times.

8) Node used to transmit the raw health data traffic from clinic to processing server.

$$\sum_{s \in CL} \sum_{d \in FN} \sum_{i \in N: i \neq j} P_{ij}^{sd} \geq \zeta a_j \quad ; \;\; \forall j \in N \tag{3-42}$$

$$\sum_{s \in CL} \sum_{d \in FN} \sum_{i \in N: i \neq j} P_{ij}^{sd} \leq M \, \zeta a_j \quad ; \;\; \forall j \in N \tag{3-43}$$

Constraints (3-42) and (3-43) ensure that $\zeta a_j = 1$ if the raw health data traffic traverses at nodes i to send the data from clinic s to the processing server at node d, otherwise it is zero.

9) Node used to transmit the analysed health data feedback traffic from processing server to clinic.

$$\sum_{s\in FN}\sum_{d\in CL}\sum_{j\in Nm[i]:i\neq j} F_{ij}^{sd} \geq \zeta b_i \quad ; \ \forall \mathrm{i} \in N \tag{3-44}$$

$$\sum_{s\in FN}\sum_{d\in CL}\sum_{j\in Nm[i]:i\neq j} F_{ij}^{sd} \leq M\ \zeta b_i \quad ; \ \forall \mathrm{i} \in N \tag{3-45}$$

Constraints (3-44) and (3-45) ensure $\zeta b_i = 1$ if the analysed health data feedback traffic traverses node i to send the analysed data from processing servers at node s to clinics d, otherwise it is zero.

10) Node used to transmit the analysed health data storage traffic from processing server to cloud storage.

$$\sum_{s\in FN}\sum_{d\in CST}\sum_{j\in Nm[i[:i\neq j} S_{ij}^{sd} \geq \theta c_i \quad ; \ \forall i \in N \tag{3-46}$$

$$\sum_{s\in FN}\sum_{d\in CST}\sum_{j\in Nm[i]:i\neq j} S_{ij}^{sd} \leq M\ \theta c_i \quad ; \ \forall i \in N \tag{3-47}$$

$$\sum_{s\in FN}\sum_{d\in CST}\sum_{i\in Nm[j]:i\neq j} S_{ij}^{sd} \geq \vartheta c_j \quad ; \ \forall j \in N \tag{3-48}$$

$$\sum_{s\in FN}\sum_{d\in CST}\sum_{i\in Nm[j]:i\neq j} S_{ij}^{sd} \leq M\ \vartheta c_j \quad ; \ \forall j \in N \tag{3-49}$$

$$\theta c_i + \vartheta c_i \; = 2\,\zeta c_i - \nu_i \quad ; \; \forall i \in N \qquad \text{(3-50)}$$

Constraint (3-46)-(3-47) ensure that $\theta c_i = 1$ if the analysed health data storage traffic traverses node i to send the analysed data from processing servers at node s to cloud storage d, otherwise it is zero. However, this does not include the last node (i.e. cloud storage) that performs the storage task. Hence, constraints (3-48)-(3-49) are to ensure $\vartheta c_j \; = 1$ if the traffic traverse node j (including the last node) while constraint (3-50) is used to determine the activation of all nodes to relay and store the analysed health data storage traffic by ensuring that the $\zeta c_i = 1$ if at least any of θc_i and ϑc_i are equal to 1 ($\theta c_i \; \mathrm{OR} \; \vartheta c_i$), otherwise zero. We achieve this by introducing a binary variable ν_i which is only equal to 1 if θc_i and ϑc_i are exclusively equal to 1 $(\theta c_i \; \mathrm{XOR} \; \vartheta c_i)$ otherwise it is zero.

11) Number of physical resource blocks used at each BS to send the raw health data traffic from clinic to processing server.

$$Pa_{ij} = \sum_{s \in CL} \sum_{d \in FN: s \neq d} \frac{P_{ij}^{sd}}{\delta a} \quad ; \; \forall i \in CL, \forall j \in BS: i \neq j \qquad \text{(3-51)}$$

$$\sum_{j \in BS} Pa_{ij} = Pt_i \quad ; \; \forall i \in CL \qquad \text{(3-52)}$$

$$\beta a_j = \sum_{i \in CL} Pa_{ij} \; Ra \quad ; \; \forall j \in BS \qquad \text{(3-53)}$$

$$\beta a_j \leq \mathrm{R} \quad ; \; \forall j \in BS \qquad \text{(3-54)}$$

Constraint (3-51) is used to ensure that each patient in the clinic is served by single BS to perform the processing task based on the traffic traversing the BS, P_{ij}^{sd}, and the size of raw health data traffic of each patient, δa, while constraint (3-52) is used to ensure that all patients are served by the BSs. Constraint (3-53) calculates the total number of PRBs used at each BS. Meanwhile, constraint (3-54) is used to ensure that the number of PRBs in each BS j do not exceed their maximum number of PRBs, Ɍ, that are dedicated for healthcare applications to perform the processing task.

12) Number of physical resource blocks used at each BS to send the analysed health data feedback traffic from processing server to clinic.

$$Pb_{ij} = \sum_{s \in FN} \sum_{d \in CL: s \neq d} \frac{F_{ij}^{sd}}{\delta b} \quad ; \ \forall i \in BS, \forall j \in CL \tag{3-55}$$

$$\sum_{i \in BS} Pb_{ij} = Pt_j \quad ; \ \forall j \in CL \tag{3-56}$$

$$\beta b_i = \sum_{j \in CL} Pb_{ij} \ Rb \quad ; \ \forall i \in BS \tag{3-57}$$

$$\beta b_i \leq \text{Ɍ} \quad ; \ \forall i \in BS \tag{3-58}$$

Constraint (3-55) ensures the analysed health data of each patient transmitted to the clinics is relayed by single BS to perform the feedback task based on the traffic traversing the BS, F_{ij}^{sd}, and the size of analysed health data feedback traffic of each patient, δb, while constraint (3-56) ensures all patients are served by the BSs. Constraint (3-57) calculates the total number of PRBs used at each BS. Constraint (3-58) is used to ensure that the number of PRBs in each BS i does not exceed its maximum number of PRBs, R, that are dedicated for healthcare applications to perform the feedback task.

13) Maximum number of patients served at each processing server.

$$\sum_{s \in CL} \omega_{sd} \leq \Omega max\ \phi_d \quad ; \ \forall d \in FN \tag{3-59}$$

Constraint (3-59) ensures that the total number of patients served by each processing server at node d, does not exceed its maximum number of users, Ωmax. However, the model also allows more than one processing server, ϕ_d, to be deployed at the same node d if the number of users is higher than Ωmax.

14) Processing and analysis time at each processing server.

$$\tau p_d = \sum_{s \in CL} m\ \omega_{sd} + \acute{c}\ \phi_d \quad ; \forall d \in FN \tag{3-60}$$

Constraint (3-60) calculates the processing and analysis time at each processing server at node d. This is based on the total number of patients served by the processing server, ω_{sd} and the number of processing servers used, ϕ_d, where m is the gradient of the graph while $\acute{c}$ is the y-intercept of the graph.

15) Storage capacity constraint at each processing server.

$$\sum_{s \in CL} \omega_{sd}\, \alpha \le \Lambda max\, \phi_d \quad ; \ \forall d \in FN \tag{3-61}$$

Constraint (3-61) ensures that the storage capacity of each processing server at node d, does not exceed its maximum capacity, Λmax. However, the model also allows more than one processing server, ϕ_d, to be deployed at the same node d if the size of the analysed data is higher than Λmax. Furthermore, this work omitted the capacity of cloud storage as a constraint, mainly because the storage capacity at the central cloud is large enough to sufficiently accommodate substantial amounts of data.

3.6 Mathematical model for the Energy efficient cloud computing health monitoring applications with LTE-M (EECC)

This section presents the MILP model that has been developed for the conventional approach (CA) to minimise the energy consumption of networking and processing without optimising the location of processing servers as the location FN is fixed at the cloud (i.e. cloud switch). Their performance will be used as a benchmark to evaluate the performance of the FOA models in terms of both energy consumption of networking equipment and processing. The same parameters, variables and objective function in Section 3.5 have been considered. However, as the location of the processing servers is at the cloud, therefore, a set of additional variables in Table 3.2 (also can be found in Appendix 1) are utilised to consider the energy consumed by the devices at the metro network, core network and cloud due to relaying the raw health data traffic and analysed health data feedback traffic.

Table 3.2: Additional variables used in EECC model

Set	
FN	Set of candidate locations to deploy PS ($FN \in CLS$)
Variables	
$ECASP$	Energy consumption of centre aggregation switches required to relay raw health data traffic
$ECASF$	Energy consumption of centre aggregation switches required to relay analysed health data feedback traffic
$EARP$	Energy consumption of aggregation routers required to relay raw health data traffic
$EARF$	Energy consumption of aggregation routers required to relay analysed health data feedback traffic
$ECRP$	Energy consumption of core routers required to relay raw health data traffic
$ECRF$	Energy consumption of core routers required to relay analysed health data feedback traffic
$ECLRP$	Energy consumption of cloud routers required to relay raw health data traffic

$ECLRF$	Energy consumption of cloud routers required to relay analysed health data feedback traffic
$ECLSP$	Energy consumption of cloud switches required to relay raw health data traffic
$ECLSF$	Energy consumption of cloud switches required to relay analysed health data feedback traffic
$ECSN$	Energy consumption of cloud server node

The energy consumption of access network, EAN, is the same as in Equation (3-3). The energy consumption of metro network, EMN, in Equation (3-15) is redefined as below:

$$EMN = (ECASP + ECASF + ECASS + EARP + EARF + EARS)\,\eta \tag{3-62}$$

where $ECASS$ and $EARS$ are the same as in Equation (3-16) and Equation (3-17), respectively, while others are given as:

$$ECASP = \sum_{i \in CAS} \left(ICAS\ x\ \zeta a_i + P_i\ \frac{PCAS - ICAS}{CCAS} \right) \tau a \tag{3-63}$$

$$ECASF = \sum_{i \in CAS} \left(ICAS\ x\ \zeta b_i + F_i\ \frac{PCAS - ICAS}{CCAS} \right) \tau b \tag{3-64}$$

$$EARP = \sum_{i \in AR} \left(IAR\ x\ \zeta a_i + P_i\ \frac{PAR - IAR}{CAR} \right) \tau a \tag{3-65}$$

$$EARF = \sum_{i \in AR} \left(IAR\ x\ \zeta b_i + F_i\ \frac{PAR - IAR}{CAR} \right) \tau b \tag{3-66}$$

The energy consumption of centre aggregation switches and aggregation routers are proportional to the size of traffic traversing them and the utilisation time. Equations (3-63) and (3-64) depict the energy consumed by the centre aggregation switches to relay raw health data traffic, $ECASP,$ and analysed health data feedback traffic, $ECASF$, respectively. Meanwhile, Equations (3-65) and (3-66) depict the energy consumed by the aggregation routers to relay raw health data traffic, $EARP,$ and analysed health data feedback traffic, $EARF$, respectively.

The energy consumption of core network, $ECN,$ in Equation (3-18) is redefined as below:

$$ECN = (ECRP + ECRF + ECRS)\,\eta \tag{3-67}$$

where the $ECRS$ is the same as in Equation (3-19) while others are given as:

$$ECRP = \sum_{i \in CR} \left(ICR\ x\ \zeta a_i + P_i\ \frac{PCR - ICR}{CCR} \right) \tau a \tag{3-68}$$

$$ECRF = \sum_{i \in CR} \left(ICR\ x\ \zeta b_i + F_i\ \frac{PCR - ICR}{CCR} \right) \tau b \tag{3-69}$$

The energy consumption of core routers is proportional to the size of traffic traversing them and the utilisation time. Equations (3-68) and (3-69) depict the energy consumed by the core routers to relay the raw health data traffic, $ECRP,$ and analysed health data feedback traffic, $ECRF$, respectively.

The energy consumption of cloud in Equation (3-20) is redefined as below:

$$ECL = (ECLRP + ECLRF + ECLRS + ECLSP + ECLSF + ECLSS \\ + ECSS + ECSTS)\ c \tag{3-70}$$

where $ECLRS$, $ECLSS$, $ECSS$ and $ECSTS$ are the same as in Equation (3-21)-(3-24), respectively, while others are given as:

$$ECLRP = \sum_{i \in CLR} \left(ICLR\ x\ \zeta a_i\ + P_i\ \frac{PCLR - ICLR}{CCLR} \right)\ \tau a \tag{3-71}$$

$$ECLRF\ = \sum_{i \in CLR} \left(ICLR\ x\ \zeta b_i + F_i\ \frac{PCLR - ICLR}{CCLR} \right)\ \tau b \tag{3-72}$$

$$ECLSP = 2 \sum_{i \in CLS} \left(ICLS\ x\ \zeta a_i + P_i\ \frac{PCLS - ICLS}{CCLS} \right)\ \tau a \tag{3-73}$$

$$ECLSF = 2 \sum_{i \in CLS} \left(ICLS\ x\ \zeta b_i + F_i\ \frac{PCLS - ICLS}{CCLS} \right)\ \tau b \tag{3-74}$$

The energy consumption of cloud routers and cloud switches are proportional to the size of traffic traversing them and the utilisation time. Equations (3-71) and (3-72) depict the energy consumed by the cloud routers to relay the raw health data traffic, $ECLRP$, and analysed health data feedback traffic, $ECLRF,$

respectively. Meanwhile, Equations (3-73) and (3-74) depict the energy consumed by the cloud switches to relay the raw health data traffic, $ECLSP$ and analysed health data feedback traffic, $ECLSF,$ respectively. Note that, the energy consumption of cloud switches to transmit the traffic is multiplied by ‘2’ for redundancy purposes [105].

The energy consumption of processing (i.e. fog node) in Equation (3-25) is redefined as below:

$$ECSN = EPS\ c \tag{3-75}$$

where EPS is the same as in Equation (3-26).

3.7 Summary

This chapter proposed a health monitoring system at the fog layer and incorporated a GPON architecture in the fog network for health monitoring applications. Two layers; fog layer and cloud layer have been proposed for the health monitoring system where the processing of the health data, the decision on the actions to treat the patients, and the temporary storage are performed at the fog layer. Meanwhile, the cloud layer is used to perform permanent storage of the processed health data. The proposed architecture of the health monitoring application consists of three main layers; the access layer where the fog resides, the metro layer to aggregate data from the access layer to the upper layer, and the core layer mainly used to permanently store the proposed analysed health data. In addition, the power profile for all equipment (i.e. network and PS), the idle power of shared networking equipment contributed by the healthcare application, and the link capacities dedicated for healthcare applications in the network are also explained in detail. Besides, the system flow of the conventional approach (CA) and the proposed fog approach (FOA) are also explained in this chapter to differentiate between those two approaches. A Mixed Integer Linear Programming (MILP) model was also presented in this chapter and used to optimise the proposed FOA (EEFC). It was used to optimise the location of PS at the access layer and for CA, it was used to optimise the processing performed at the central cloud.

Chapter 4

Energy efficient fog computing with Long Term Evolution for machine (LTE-M) for ECG monitoring applications

4.1 Introduction

In this chapter, we investigate the use of fog computing for health monitoring applications considering a sample realistic dataset where we considered, as respondents, outpatients at West Leeds, United Kingdom, who suffered from cardiovascular disease (CVD) and underwent cardiac surgery. Respondents with CVD diseases were selected mainly because CVD has emerged as the top cause for mortality worldwide and is expected to reach 23.3 million by 2030 [106], [107]. Precisely, this study monitored postoperative atrial fibrillation (AF), a common cardiac events following cardiac surgery [108]. A total of 37 clinics located at West Leeds were selected to monitor patients with postoperative AF in cardiac surgery. The total number of patients from each clinic was applied to calculate the number of patients who experienced postoperative AF, which reflected the traffic demands in this study. As 92% of the patients registered to the clinics resided within 2 km [109], this study monitored the patients from their homes or outside the clinics.

Among all the available medical monitoring services, the ECG analysis happens to be the most common clinical cardiac test [106], [110]. Thus, fog computing was incorporated with the network edge to carry out ECG feature extraction and analysis. This work also demonstrates edge fog computing effectiveness in terms of energy consumption for both networking equipment and processing. In this work, a Mixed Integer Linear Programming (MILP) model introduced in Chapter 3 is used to optimise the number and location of processing servers at the network edge so that the energy consumption of both networking equipment and processing are minimised. A heuristic model, based on the insights from the results obtained from the MILP model, is also developed for real-time implementation.

4.2 Parameters consideration

This section elaborates in detail the methodologies used for determining the model input parameters considered in this chapter. The input parameters are divided into several parts such as network layout under GPON network, the number of monitored patients in West Leeds, UK, the processing time of health data (i.e. ECG signal) and the calculation of data rate for traffic transmission.

4.2.1 Network layout under GPON network in West Leeds, UK

In this chapter, the West Leeds area was considered as a case study to examine the energy efficiency of fog computing for health monitoring applications. The patients were considered to be located in the clinics (i.e. within LTE-M base station coverage around the clinic) due to the uncertainty of their precise locations. A total of 37 clinics were available in West Leeds in 2014 / 2015, see Figure 4.1, [111]. While all 37 clinics had been considered in this work, only those that appeared to be close to BSs, i.e. potential BSs to serve patients were selected for further analyses by looking into the distance between the clinics and the BSs. Note that the locations of clinics and BSs (i.e. latitude and longitude) refer to the actual locations found in West Leeds, which had been obtained from Google Maps based on the names of clinics listed by [111] in 2014 / 2015 and OFCOM UK Mobile Site finder published in May 2012 [112], respectively. With that, the distances between the clinics and BSs were determined based on their latitudes and longitudes. In this work, LTE-M was opted to serve the health application with a coverage radius less than 11 km [113]. Hence, patients could be served by a BS within 11 km from their registered clinics. As for this work, 315 BSs were considered as they were located less than 11 km from any clinic.

The OLT was also deployed in the network based on the location of a local exchange provided by BT Wholesale network [114]. The distances between the considered BSs and the OLTs were calculated based on latitude and longitude. Furthermore, in the GPON network, the BSs are co-located with the ONUs and the maximum allowed distance from ONU to OLT is 20 km due to optical signal integrity considerations [115], [116]. With that, only 88 OLTs were selected as they are located within 20 km from the ONUs co-located with the BSs. Figure 4.1 illustrates the locations distribution for the 37 clinics, the 315 BSs/ONUs and the 88 OLTs, respectively, in West Leeds, UK. This particular area had been considered as it covers all 37 clinics.

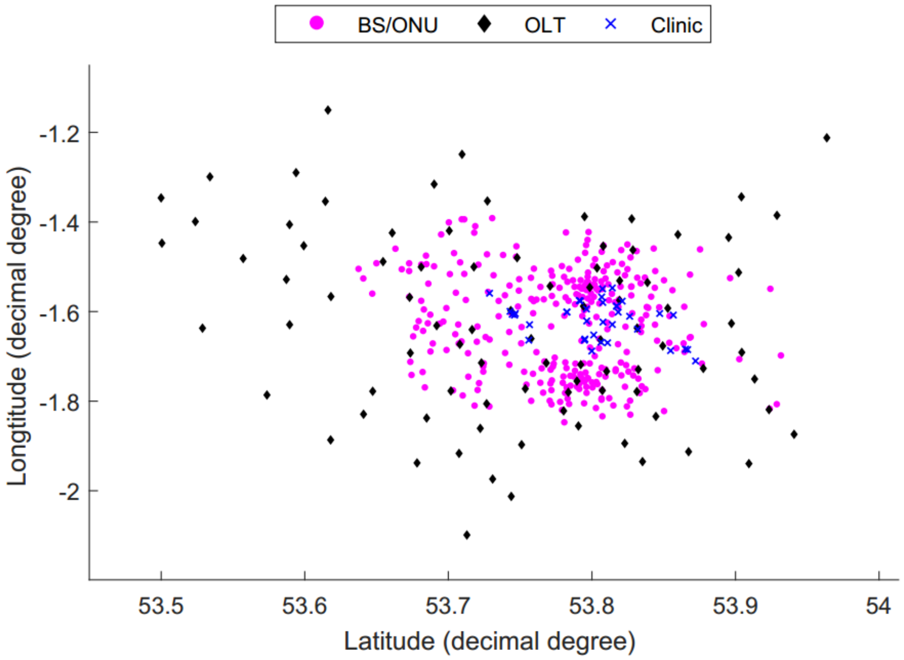


Figure 4.1: BS, OLT and clinic locations in West Leeds

However, due to the limitation of MILP to run the model with vast number of nodes, the number of BSs was reduced to evaluate the conventional approach (CA) and fog optimised approach (FOA) as explained in Section 3.4. Besides, with the shortcoming in M2M device in terms of its limited available power and the need to reduce consumption, the devices were connected to the nearest BS [117]. As such, only 26 BSs were considered in this study as they appeared to be the nearest BSs to the clinics that served patients. A reduction in the number of BSs embedded within the network decreased the number of OLTs to 75, in which they are within 20 km from the 26 BSs. Then, we optimised the network connection at the access layer (i.e. connections between ONUs and OLTs) within the GPON network using MATLAB. The link between ONU and OLT within the GPON network is only legitimate if the distance is equal to or less than 20 km. In GPON networks, an average of 10 to 30 LTE base stations are connected to a single PON [118]. The present

GPON technologies utilise splitters with split ratios of 1:4, 1:8, 1:16, 1:64 and 1:128. The ratio 1:16 was selected for implementation in the network studied. In order to provide resilience, each splitter is connected to 2 PONs in the OLT which has 4 PONs, hence each OLT can support up to 32 LTE BSs/ONUs. Due to this, we only consider one OLT to be connected to the 26 ONUs (co-located with the BS) within the network. Note that the location of OLT is selected by considering the lowest distances with the 26 BSs/ONUs. Figure 4.2 presents the network layout of the GPON network after optimisation, where the black diamond reflects the optimal OLT selected in the network.

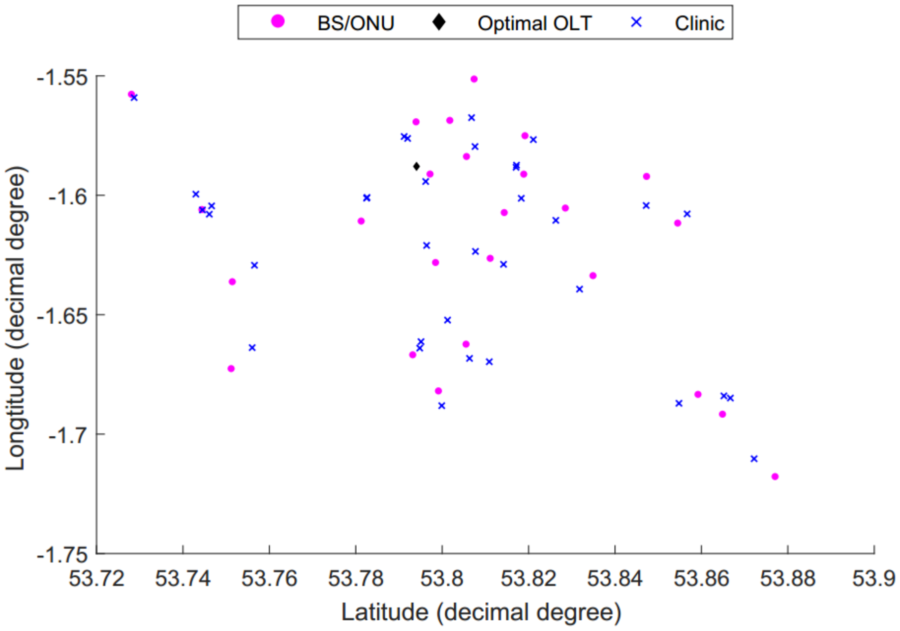


Figure 4.2: Selected BSs and OLT to serve clinics in West Leeds

4.2.2 Total number of monitored patients in West Leeds, UK

According to the British Heart Foundation, the total UK population of those aged 18 years old and above suffering from Coronary Artery Bypass Surgery

(CABG) and Percutaneous Coronary Interventions Surgery (PCIs) that include surgeries performed in NHS and selected private hospitals in 2014 are 17,513 and 96,143, respectively [119]. The Office for National Statistics (ONS) has further claimed that the UK population aged 18 years and above in 2014 is 80% of the total population [120]. Based on the collected and calculated figures, the percentage of patients from the UK population that had undergone heart surgeries (CABGs and PCIs) in 2014 is 0.22%. To estimate the number of monitored patients that may experience postoperative AF in West Leeds, UK, 0.176% of the total number of patients registered in 37 clinics in West Leeds [111] were selected as an upper limit to reflect the traffic demands in the network. Table 4.1 presents the deduced total number of patients registered at each clinic who are expected to experience postoperative AF.

Table 4.1: Number of monitored patients in clinics

Clinic	Number of Patients	Clinic	Number of Patients
Craven Road Medical Practice	20	Leeds Student Practice	68
Hyde Park Surgery	18	Burton Croft Surgery	20
Laurel Bank Surgery	13	Kirkstall Lane Medical Centre	15
Burley Park Medical Centre	23	Thornton Medical Centre	16
Gildersome Health Centre	6	The Dekeyser Group Practice	30
Leigh View Medical Practice	29	West Lodge Surgery	32
Hillfoot Surgery	13	Dr. KW McGechaen & Partner	8
Pudsey Health Centre	13	Robin Lane Medical Centre	24
Dr. S M Chen & Partner	8	Beech Tree Medical Centre	4
Hawthorn Surgery	10	Priory View Medical Centre	16

High Field Surgery	14	Abbey Grange Medical Centre	16
Vesper Road Surgery	11	Fieldhead Surgery	10
Manor Park Surgery	27	The Highfield Medical Centre	9
Dr. G Leeds & Partners	25	Dr. F Gupta’s Practice	6
Guiseley and Yeadon Medical Practice	21	Park Road & Menston	19
Yeadon Tarn Medical Practice	12	Rawdon Surgery	14
Dr. KJ Manock & Partners	44	Whitehall Surgery	16
Dr. JA Browne’s Practice	28	Dr. N Saddiq’s Practice	5
Dr. JJ McPeakes Practice	6		

4.2.3 Time measurement for processing and analysis of Electrocardiogram (ECG) signal using Pan-Tompkins algorithm

The ECG signals, which are required to monitor postoperative AF among cardiac surgical patients, were based on the MIT_BIT Arrhythmia database [121], [122]. Although 30 minutes of ECG recording was provided, we only consider 30 seconds, as illustrated in Figure 4.3. Note that, the 30-second ECG signal offers accurate results for the analysis, as recommended in [108], and such 30 seconds of un-processed ECG signals have a volume of 252.8 kbits. The ECG signals were processed using the Pan-Tompkins algorithm, which is a resource-demanding algorithm [106] with 99.3% accuracy [110], to extract heart rate and QRS duration [123], [124] for further analysis. The calculation of the heart rate from the 30-second ECG signal is based on the number of R waves within the 30 seconds and this number was multiplied by 2 to obtain the heart rate in beats per minute [124], while the QRS duration was obtained based on the time between Q and S waves found in the ECG signal [124], [125].

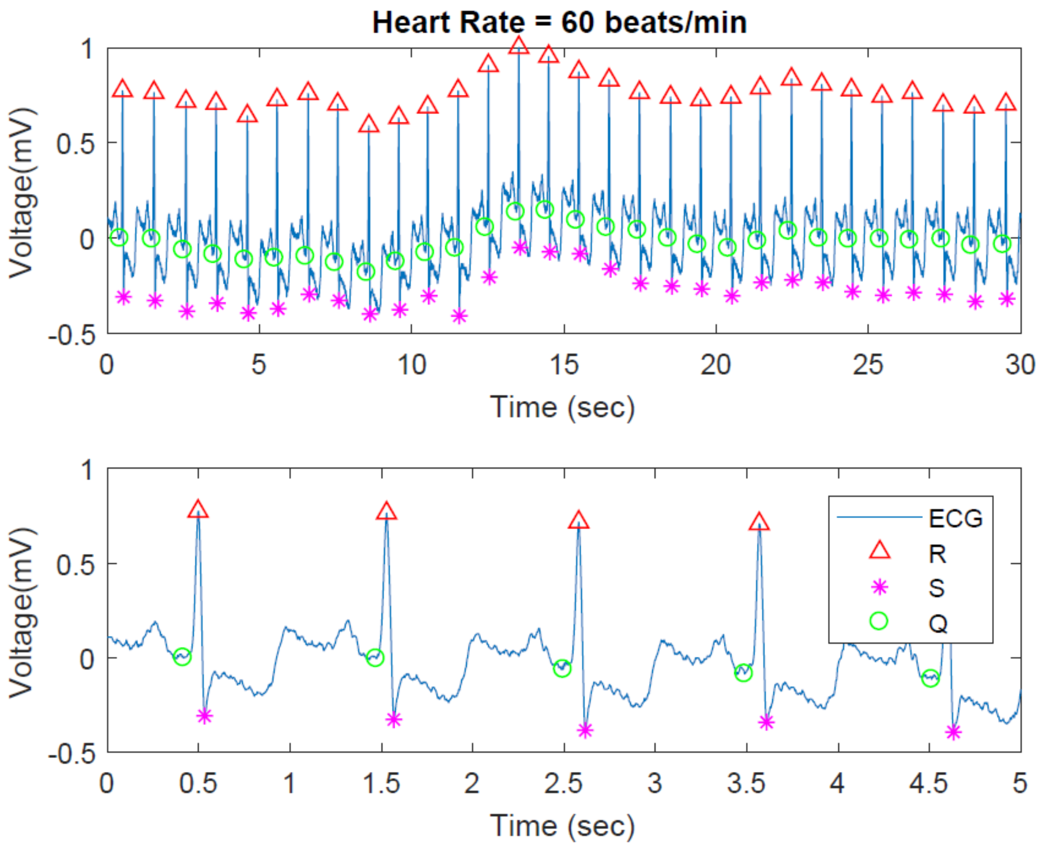


Figure 4.3: The 30-second and 5-second ECG waveform

The processing server selected in both fog and central cloud is Intel Core i5-4460 with 3.2 GHz CPU and 500 GByte hard drive [126]. An experiment was conducted using MATLAB with a parallel processing function to determine the correlation between time and number of patients for processing and analysis of raw ECG data using Pan-Tompkins algorithm. This was carried out by performing the processing task on the 30-second ECG signals generated by 10k to 50k patients in 10k steps. At each 10k step, the processing operation was repeated 5 times to calculate the average time for the processing duration. Note that, the 30-second ECG signals are made up of 1 ECG record repeated for all patients. Also, note that the time to perform the processing using MATLAB consists of both the time to submit the data for parallel processing and the time to run the algorithm. The results were then fitted with a linear line (dotted blue line), as illustrated in Figure 4.4. For instance, a 10-second duration for processing could cover 2657 patients. We also obtained the correlation between the time and number of patients for the processing and analysis of raw ECG signal considering 41 ECG records retrieved from the MIT_BIT Arrhythmia database [121], [122] with a duration of 30-seconds each. Note that the 41 ECG records are repeated to cover all patients. For instance, a maximum of 244 patients are represented by the same ECG recording when the total patients are 10k. The results are as shown as a red line in Figure 4.4. The two experiments with a single ECG signal and multiple ECG signals have resulted in similar linear relationships. The time to process and analyse the raw ECG signal is determined considering the single ECG signal.

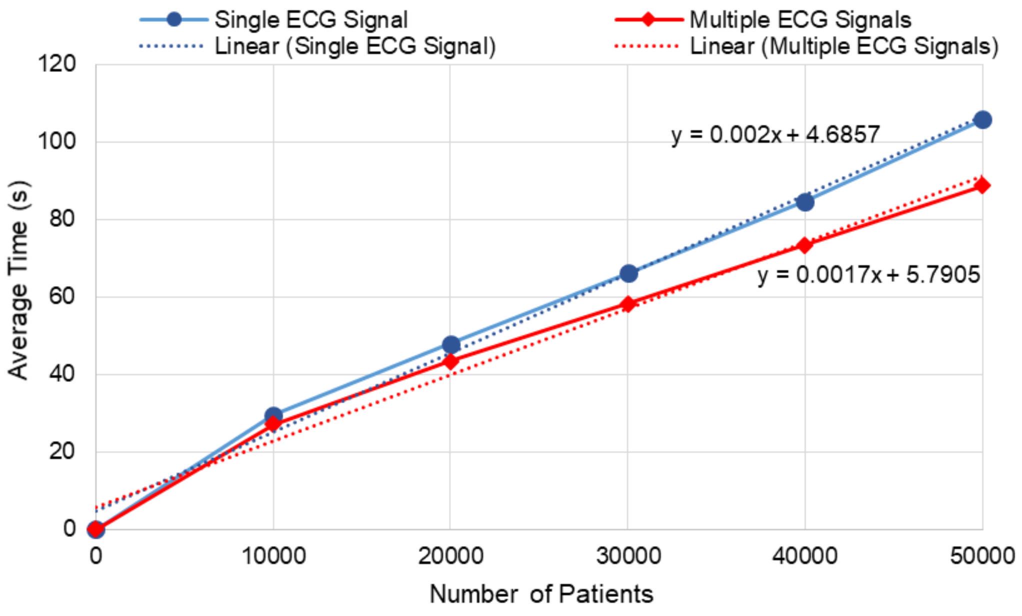


Figure 4.4: Number of patients versus time, based on MATLAB simulations

In addition, based from the experiment performed for processing and analysis of the raw ECG data, the τp_d in Equation (3-60) is obtained from the total number of patients served by processing server at node d, as given in Equation (4-1):

$$\tau p_d = \sum_{s \in CL} m \, \omega_{sd} + \acute{c} \, \phi_d \quad ; \forall d \in FN \tag{4-1}$$

where, m and $\acute{c}$ equal 0.002 and 4.6857, respectively.

4.2.4 Data rate calculation for traffic transmission in the network

The American Heart Association (AHA) has recommended that the golden time to save a heart patient's life by sending an alarm message to a cardiologist upon detection of a sudden fall or rise in cardiac vital signs is between 4 and 6 minutes [57]. As such, 4 minutes, τt, was selected for this work as the maximum duration imposed by AHA to calculate the minimum data rate for each patient. Note that this 4-minute duration should include the following: i) the time to record the 30-second ECG signal, τm, ii) the time to transmit ECG signals to the processing server for the processing task, τmax, iii) the time for processing and analysis, τp and iv) the time to transmit the analysed ECG data for feedback, τb as illustrated in Figure 4.5. Therefore, latency is not considered in this work as the time to perform the main tasks explained above to save the heart patents is limited to 4 minutes. Note that Figure 4.5 also includes the time to transmit the analysed ECG data to the cloud for permanent storage, τc. Also note that any propagation delay in this work is dismissed as the time is too short (milliseconds).

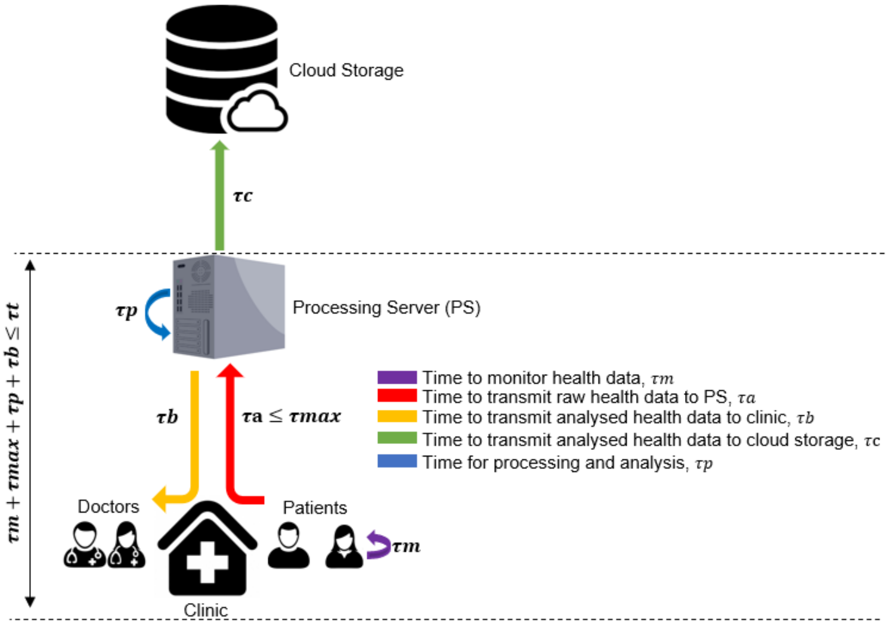

Cloud Storage
τc
Processing Server (PS)
τp
τb
$\tau a \leq \tau max$
$\tau m + \tau max + \tau p + \tau b \leq \tau t$
🟪 Time to monitor health data, τm
🟥 Time to transmit raw health data to PS, τa
🟨 Time to transmit analysed health data to clinic, τb
🟩 Time to transmit analysed health data to cloud storage, τc
🟦 Time for processing and analysis, τp
Doctors
Patients
τm
Clinic

Figure 4.5: Transmission times for each task and processing time

In this study, we considered all monitored patients (669 patients) were served by the same processing server. Due to this, the time to process 669 patients using the same processing server, τp, based on the fitting in Figure 4.4 is 6.02 seconds. The time to perform the feedback transmission, τb is based on the provisioned data rate for each user and the size of the analysed data. To determine the data rate for feedback in FOA, the minimum shared link capacity at the edge network provisioned for healthcare application where the processing server is located (i.e. the link between the ONU and OLT) is considered. By considering all patients to be served by the same processing server, this minimum link capacity was divided equally among the patients, hence giving a data rate, δb, of 350 bps for each patient in FOA. However, the minimum link capacity in CA is the downlink capacity between the OLT and the ONU as the processing server is located at the central cloud, therefore, a data rate of 700 bps for each patient is considered in CA.

Without a custom communication system, this particular data rate may not be supported. Therefore, the LTE-M that is exclusively designed for M2M applications was employed. The LTE-M operates on a 1.4 MHz carrier where 6 resource blocks (RBs) are equipped for a duration of one time slot (0.5 ms) [113], [127]. In LTE, the smallest modulation structure is a resource element (RE) which has one subcarrier of 15 kHz by one symbol [127], [128]. The resource elements are grouped into a resource block (RB) with 12 subcarriers for a duration of one slot (6 or 7 symbols) with a 180 kHz bandwidth as illustrated in Figure 4.6. Therefore, in one slot for 1 RB, there are 84 REs (i.e. 12 subcarriers x 7 symbols).

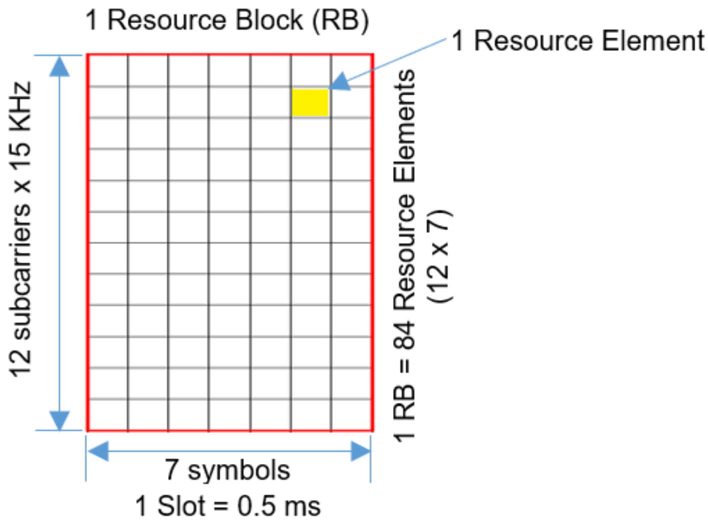


Figure 4.6: LTE-M resource grid

For a 1 second duration, a maximum of 12,000 RBs are available which supports a total data rate of 2.016 Mbps with 168 bps per single RB when using Quadrature Phase Shift Keying (QPSK) as the modulation format per RE. Meanwhile, the transmission time interval (TTI) in LTE is 1 ms which is 1 subframe (i.e. 2 slots) [129], hence a minimum of 2 RBs (i.e. 1 physical resource block, PRB) with a data rate of 336 bps can be scheduled for each M2M device. Due to this, we allocate 1 PRB to each patient which gives a data rate (δb) of 336 bps to transmit the analysed ECG data for feedback purposes in FOA, while in CA we allocate 2 PRBs for each patient which gives a data rate of 672 bps. Note that given a data rate with a value higher than 336 bps and 672 bps to each patient in FOA and CA, respectively, will exceed the link capacity (i.e. between ONU and OLT) that is provisioned for healthcare in the network. The size of the processed ECG data using the Pan-Tompkins algorithm is reduced from 252.8 kbits to 256 bits for each patient. Hence, the feedback time, τb to transmit the analysed ECG data (256 bits) for each patient with the given data rate (δb) in FOA and CA will require 0.76 s and 0.38 s, respectively (i.e. $256 \text{ bits}/\delta b$).

Note that we choose to limit the feedback data rate by data rate available for healthcare applications in the GPON links to increase the feedback data rate. Therefore, we use more resources to transmit a feedback signal (256 bits) to decrease the feedback time which, in turn, gives more time to transmit the raw ECG signal (252.8 kbits). This high available time to transmit the raw ECG signal uses a lower data rate which will result in activating fewer BSs. Note that, activating fewer BSs for a longer time is more efficient than activating a large number of BSs for a shorter time as the idle power consumption of a BS is 63% of its total power. Therefore, the remaining maximum time to send raw ECG signals to the processing server, τmax, had been set as 203.2 s and 203.6 s in FOA and CA, respectively (i.e. $\tau max = \tau t - \tau m - \tau p - \tau b$). As the size of a 30-second ECG signal is 252.8 kbits, the minimum data rate, $\delta min,$ required for each patient for FOA and CA is 1.244 kbps and 1.241 kbps (i.e. $\delta min = 252.8\ \mathrm{kbits}/\tau max$), respectively. Note that the calculated data rate refers to the minimum data rate that should be disseminated to each patient so as to ensure that the system works within the 4 minutes, as required by AHA. Nevertheless, the data rate provided to each patient relies on the type of wireless technology used in this work, which is LTE-M. As the TTI to each user/M2M device with LTE-M had been 1 PRBs (336bps), the minimum data rate that was offered to each patient, δa, in this work for all approaches was 1.344 kbps, which is equivalent to 4 PRBs per patient. As the size of raw ECG data is 252.8 kbits, the transmission time required to send the ECG data to the processing server, $\tau a,$ for both FOA and CA is 188.1s (i.e. $252.8\ \mathrm{kbits}/\delta a$).

The data rate to send the processed ECG data at the processing server to the cloud storage for permanent storage was determined by considering the lowest shared link capacity or devices capacity from the processing server to the cloud storage, $Cmin$, where the provision capacity for health M2M application are 234.4 kbps and 5.4 Mbps for FOA and CA, respectively. Furthermore, when the worst-case scenario is considered where all 669 patients shared similar physical link to send their processed ECG data to the cloud storage, the lowest link or device capacity was divided equally to each patient, hence giving a data rate, δc, of 350 bps and 8.07 Kbps for each patient for FOA and CA, respectively (i.e. $\delta c = Cmin\ /669$). As such, the time taken to send the analysed data to the cloud storage, τc, with the given data rate $(\delta \mathrm{c})$

in FOA and CA is 0.73 s and 0.032 s, respectively (i.e. $\tau c = 256/\delta c$) . Table 4.2 shows the input parameter calculated for the FOA and CA as discussed above.

Table 4.2: Parameter inputs for FOA and CA

Parameter	FOA	CA
Size of ECG data (kbits)	252.8	252.8
Size of analysed ECG data (bits)	256	256
Transmission time to transmit ECG data to processing server, τa (s)	188.1	188.1
Data rate to transmit ECG data to processing server, δa (bps)	1344	1344
Transmission time to transmit analysed ECG data to clinic, τb (s)	0.76	0.38
Data rate to transmit analysed ECG data to clinic, δb (bps)	336	672
Transmission time to transmit analysed ECG data to cloud storage, τc (s)	0.73	0.032
Data rate to transmit analysed ECG data to cloud storage, δc (bps)	350	8070

4.3 Performance evaluation for the EEFC Model

This section presents the results and analysis of the EEFC model for fog computing approach (FOA) and the EECC model for the conventional approach (CA). AMPL software with CPLEX 12.8 solver running on high-performance computing (HPC) cluster with a 12 core CPU and 64 GB RAM was used as the platform for solving the MILP models. The performance of the EECC model was used as a benchmark to evaluate the performance of the EEFC model in terms of energy consumption of both networking equipment and processing. The evaluation of the two models is performed using the GPON architecture, as shown in Figure 3.2, with 26 BSs and 1 OLT. Note that in this work the energy consumption at the IoT layer was neglected as we only considered the energy consumed by the shared networking equipment. The input parameter calculated as in Table 4.2 and the input parameter for the networking and computing devices in Table 4.3 were used to obtain the results. Also, we consider a scenario where we only allow one processing server at each candidate node (i.e. fog node) as the limited space at the fog node can be shared by multiple applications, i.e. ϕ_d will be a parameter $\phi_d = 1$.

As discussed in Section 3.3.2, the healthcare application is considered to contribute to 0.3% of the idle power of the networking devices. However, the LTE-M BS shares capacity, antenna, radio and hardware with the legacy LTE networks (20MHz) [25]. Due to this, the idle power of the BS contributed by healthcare application is calculated based on a 7% allocation of LTE-M network from the legacy LTE network (1.4MHz/20MHz) and 6% allocation of healthcare applications from the total M2M applications supported by the LTE-M network. The processing server is the most energy-consuming device in the network as the processing servers are dedicated to the healthcare application, hence maximum idle power is consumed.

Table 4.3: Input parameters for networking and computing devices

Parameter	Value
Maximum power consumption of core router (CRS-3), PCR	12300 W [130]
Core router capacity (CRS-3), CCR	4480 Gbps [130]
Maximum power consumption of cloud switch (Catalyst 6509), $PCLS$	2020 W [130]
Cloud switch capacity (Catalyst 6509), $CCLS$	320 Gbps [130]
Maximum power consumption of cloud router (7609), $PCLR$	4550 W [130]
Cloud router capacity (7609), $CCLR$	560 Gbps [130]
Maximum power consumption of content server, PCS	380.8 W [131]
Idle power consumption of content server, ICS	324.82 W [131]
Content server capacity, CCS	1.8 Gbps [131]
Maximum power consumption of cloud storage, $PCST$	4900 W [102]
Cloud storage capacity $CCST$	75.6 TB [102]
Maximum power consumption of aggregation router (7609), PAR	4550 W [13], [130]
Aggregation router capacity (7609), CAR	560 Gbps [13], [130]
Maximum power consumption of centre aggregation switch, (Catalyst 6509), $PCAS$	1766 W [130]
Centre aggregation switch capacity (Catalyst 6509), $CCAS$	256 Gbps [130]
Maximum power consumption of OLT, $POLT$	20 W [116]
OLT capacity, $COLT$	128 Gbps [116]
Maximum power consumption of ONU, $PONU$	8 W [99]
ONU capacity, $CONU$	3.75 Gbps
Maximum power consumption of LTE Base Station, PBS	528 W [132]

Idle power consumption of LTE Base Station, IBS	333 W [132]
Maximum power consumption of processing server, PPS	180 W [133]
Idle power consumption of processing server, IPS	78 W [133]
IP over WDM, access and metro network PUE, η	1.5 [102], [103]
Cloud and fog PUE, c	2.5 [105]

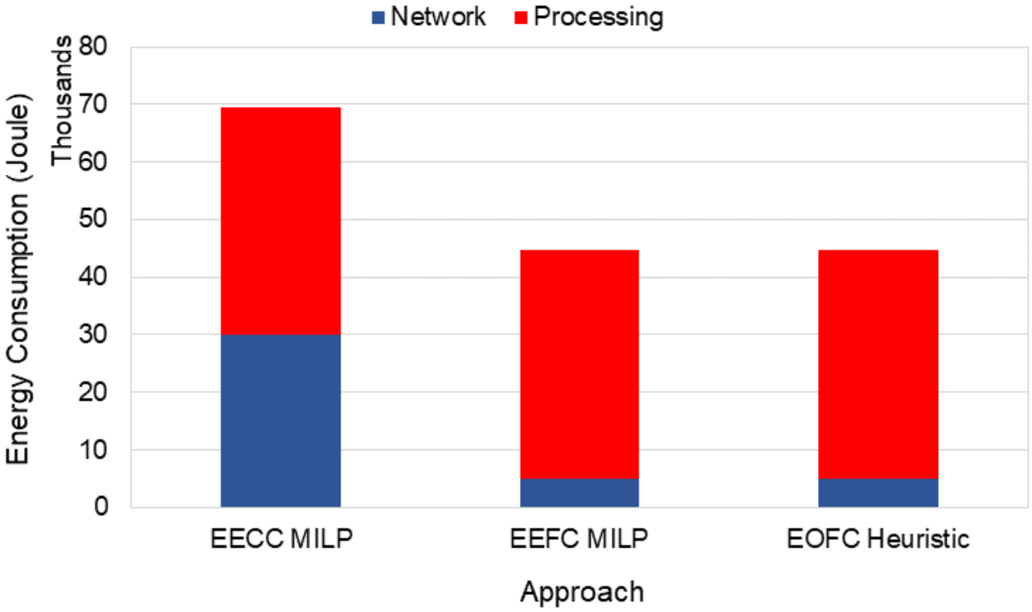


Figure 4.7: Energy consumption of networking equipment and processing in GPON architecture

Figure 4.7 shows the energy consumption of networking equipment and processing in the GPON architecture for EECC model, EEFC model and EOFC heuristic. We used the EECC model as our benchmark to evaluate the performance of energy consumption in the EEFC model. The energy saving of networking equipment in the EEFC model compared to the EECC model is 83.1%, as illustrated in Figure 4.7. This saving is due to two factors. The first is the size of data traversing the network equipment in the metro and core layers. The second factor is the duration of utilising the network equipment in the metro and core layers. Note that the bigger the size of the data, and the longer the time duration of transmission, the higher the energy consumption. Below, we explore in more details the role of the two factors above in minimising the energy consumption of networking equipment in the EEFC model compared to the EECC model.

The size of data in the EEFC model traversing from the centre aggregation switch (CAS) at the metro network to the central cloud is small compared to the un-processed data in the EECC model. This is because in the EEFC model, the location of PS is optimised at the access layer. The MILP results show that there is only one PS deployed at the OLT as it is the nearest shared point to the patients (the OLT is connected to all BSs in the network). Due to this, the aggregated ECG signals from patients in the EEFC model are processed in the fog located at OLT, which reduces the data size. This reduces the energy consumed by the networking equipment in the EEFC model when the analysed data traverse from the CAS at the metro network to the central cloud for permanent storage as the energy is partly proportional to the size of data. Comparing that to EECC model, higher energy is consumed by the networking equipment in the metro and core layers in the EECC model as the un-processed data are sent to the central cloud to be processed. Note that permanent storage is also performed in the EECC model after the data is processed.

In the EECC model, all devices in the three layers are utilised for 188.48 s ($\tau a + \tau b$) that includes the time to send the un-processed data to PS for processing and analysis, (τa), and the time to send the analysed data to the clinics (i.e. doctors) for feedback purpose, (τb), except the cloud switches, content servers and cloud storage. The content servers and cloud storage are utilised for 32 ms, (τc), to perform the storage task. The cloud switches are used three times in the network: first, to send the un-processed data to PS, then to send the analysed data to the clinics for feedback, and, finally, to send the analysed data to the cloud for permanent storage. Due to this, the cloud switches are utilised for $\tau a + \tau b + \tau c$.

In the EEFC model, the BS and ONU are utilised for 188.86 s ($\tau a + \tau b$) that include the time to send the un-processed data to PS at the OLT for processing and analysis, (τa), and the time to send the analysed data to the clinics for feedback purpose, (τb). The OLT is used three times in the network: to send the un-processed data to its co-located PS, and to send the analysed data to the clinics for feedback and to the central cloud for permanent storage, hence the OLT is utilised for $\tau a + \tau b + \tau c$. However, if the PSs are placed at the ONUs, therefore, the time of these ONUs are utilised is $\tau a + \tau b + \tau c$. Note that the time to send the analysed data to the cloud storage in the EEFC model, (τc), is 0.73 s. Meanwhile, the equipment at the metro and core layers is only used to send the analysed data to the cloud storage for permanent storage. Due to this, the utilisation time of equipment at the metro and core layers in EEFC model is, (τc), 0.73s.

Figure 4.7 illustrates that the energy consumption for processing in the EEFC model is slightly higher than the EECC model by 0.53%. This is due to the high utilisation time of the processing server in the EEFC model compared to the EECC model. Recall that, in the EEFC model, the processing server is utilised for 0.76 s and 0.73 s to send the analysed data for feedback and permanent storage, respectively, while it is 0.38 s and 32 ms in the EECC model. This is due to the link capacity or devices capacity limitation in the access layer where the processing server is located in the EEFC model which limits the data rate to send the analysed data to the clinic and cloud storage compared to the EECC model. However, the total energy saving that includes the networking equipment and processing in the EEFC model compared to the EECC model is 35.7%.

4.4 The Energy optimised fog computing (EOFC) heuristic

The Energy Optimised Fog Computing (EOFC) heuristic was developed as a method to validate the MILP operation and to deliver a real-time solution of the FOA. Compared to the state-of-the-art of heuristic algorithm developed using various techniques to achieve their objectives, the EOFC heuristic is developed based on the insights from the results obtained from the MILP model to minimise the energy consumption of the networking equipment and processing. In this section, we explain the flow of the EOFC heuristic model based on the provided flow chart. Then, we discuss the performance of EOFC heuristic model compared to the EEFC model in terms of energy consumption of both networking equipment and processing.

4.4.1 EOFC heuristic description

The heuristic determines the BSs to serve patients to send raw health data and receive feedback data; and the nodes to place processing servers at the access network so that the energy consumption of both networking and processing is minimised. Figure 4.8 shows the flow chart of the EOFC heuristic.

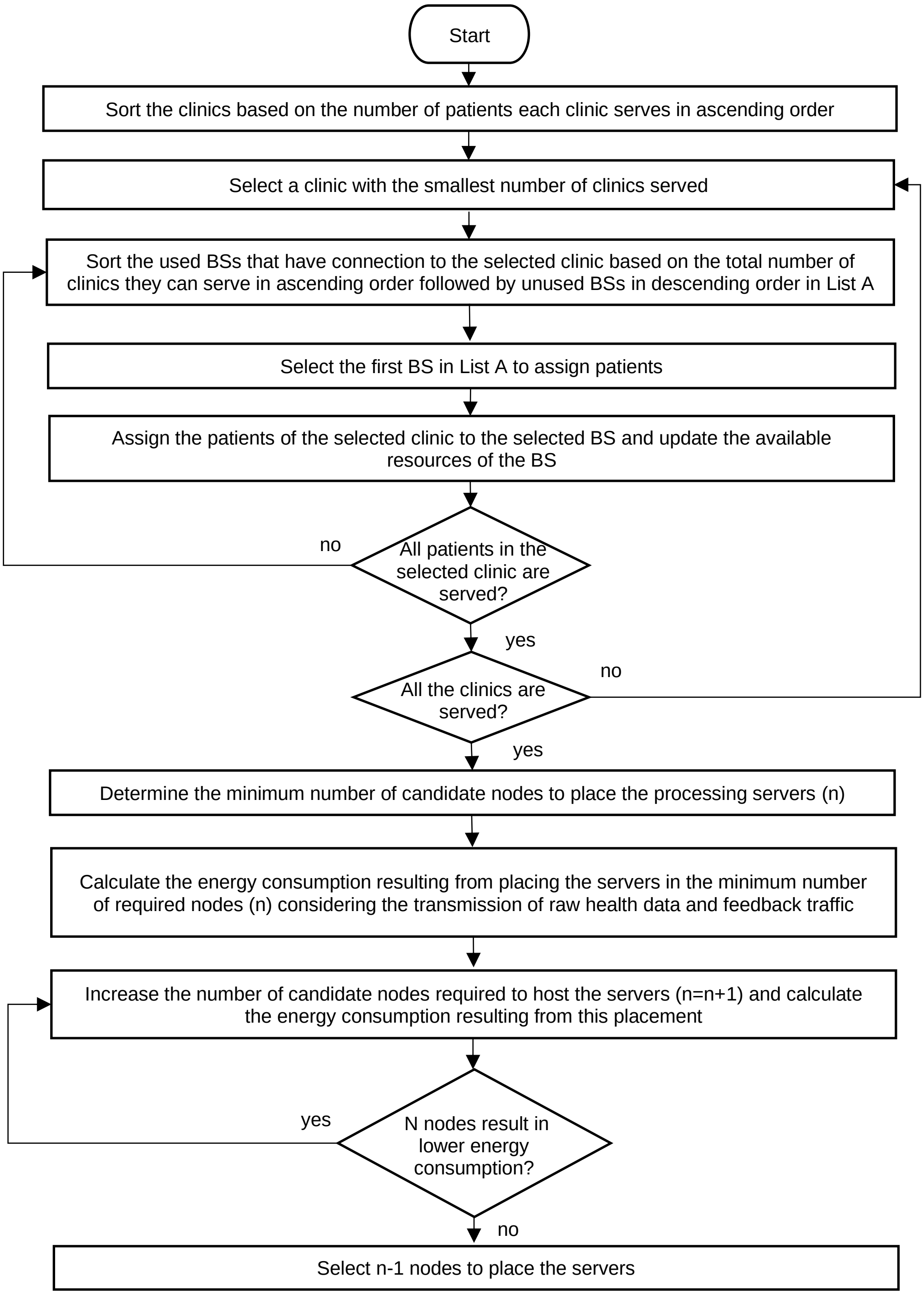

Start
Sort the clinics based on the number of patients each clinic serves in ascending order
Select a clinic with the smallest number of clinics served
Sort the used BSs that have connection to the selected clinic based on the total number of clinics they can serve in ascending order followed by unused BSs in descending order in List A
Select the first BS in List A to assign patients
Assign the patients of the selected clinic to the selected BS and update the available resources of the BS
no
All patients in the selected clinic are served?
yes
no
All the clinics are served?
yes
Determine the minimum number of candidate nodes to place the processing servers (n)
Calculate the energy consumption resulting from placing the servers in the minimum number of required nodes (n) considering the transmission of raw health data and feedback traffic
Increase the number of candidate nodes required to host the servers (n=n+1) and calculate the energy consumption resulting from this placement
yes
N nodes result in lower energy consumption?
no
Select n-1 nodes to place the servers

Figure 4.8: Flow chart for EOFC heuristic

In this process, the heuristic begins by sorting the clinics based on the number of patients the clinic serves in ascending order. The heuristic assigns first the clinic with the smallest number of patients to BSs to help in packing the BSs (packing is optimum when equipment have high idle power consumption). The assignment of clinic patients to a BS is as follows: The heuristic sorts the BSs that have a connection to the clinic under consideration starting with BSs previously used by the healthcare application that has available resources. These BSs are sorted in ascending order based on the total number of clinics the BS can serve followed by the unused BSs in descending order. Ascending order of activated BSs reduces the number of utilised BS while, the descending order of unused BSs ensures that options are left open until late in the allocation process. Then, the patients of the clinic under consideration are consolidated to the minimum number of BSs to reduce the number of BSs used by the healthcare application.

The heuristic then determines the number of processing servers required to serve the patients and the nodes hosting them. The candidate nodes that can host the servers are the ONUs connected to the BSs selected to serve the patients and the OLTs. Considering the minimum number of candidate nodes required to host servers to serve all the patients (which is based on the maximum number of servers a node can host), the heuristic finds the combination of candidate nodes to host processing servers that result in minimum energy consumption. The aim of limiting the number of candidate nodes to place the processing servers is to reduce the utilisation of the Ethernet switches to serve the processing servers.

This energy consumption that results from hosting servers at a combination of candidate nodes is calculated by routing the traffic (raw health data traffic) from BSs (starting with the BS serving the largest number of patients) to the nearest node with available processing capacity out of the combination of candidate nodes under consideration based on minimum hop routing. Also, BSs to send feedback traffic from the combination of candidate nodes to clinics are selected using the same approach used to select BSs to send raw health data. Note that BSs different from those used to send raw health data are used to send feedback traffic as the size of the analysed health data feedback traffic is smaller than the raw health data traffic. Therefore, a small number of BSs are utilised to send analysed health data feedback traffic.

The combination of nodes hosting servers considering the minimum number of candidate nodes required to host servers to serve all the patients that result in minimum energy consumption is selected.

The heuristic increases the number of candidate nodes to host servers and repeats the above process. The energy consumption resulting from using this combination of nodes is calculated and compared to the energy consumption resulting from the combination of nodes hosting servers considering the minimum number of candidate nodes required to host servers. If the latter is lower, the heuristic examines placing servers in more candidate nodes. If the former is lower, the minimum number of candidate nodes required to host servers is selected to place servers.

4.4.2 Performance evaluation of the EOFC heuristic

In this section, the performance of the Energy Optimised Fog Computing (EOFC) heuristic compared to the Energy Efficient Fog Computing Health Monitoring Applications with LTE-M (EEFC) model is evaluated. As shown in Figure 4.7, the EOFC heuristic has the same performance as the EEFC model in terms of the network energy consumption which gives the same energy saving of 83.1%, in comparison to Energy Efficient Cloud Computing Health Monitoring Applications with LTE-M (EECC) model. This is because the optimal location to place the processing server in both EOFC heuristic and EEFC model is at the OLT, and the same amount of networking equipment is utilised to serve the patients. The figure also illustrates that the EOFC heuristic has the same performance as the EEFC model in terms of the processing energy consumption, which gives equal energy increase, in comparison to the EECC model. We also evaluate the computational time to run the EOFC heuristic and the EEFC model. The results show that the EOFC heuristic running on a normal PC with 3.2 GHz CPU and 16 GB RAM took 11.1 sec to finish while the EEFC model running on high-performance computing (HPC) cluster with a 12 core CPU and 64 GB RAM was manually stopped after 48 hours.

4.5 Impact of idle power of networking and processing equipment on the EEFC model under GPON access network

In this section, the impact of the idle power of networking and processing equipment on the energy efficiency of the EEFC model and the EECC model under GPON access network is evaluated by reducing the idle power of all equipment by 30% and 60% from its maximum power. Note that as the idle power of the base station, processing server and content server are obtained from data sheets and references in [134], [133], and [131], respectively, and not generalised to be 90% of the maximum power, and to obtain an equivalent reduction ratio for all equipment, we considered reductions by 33% and 67% from their fixed idle power.

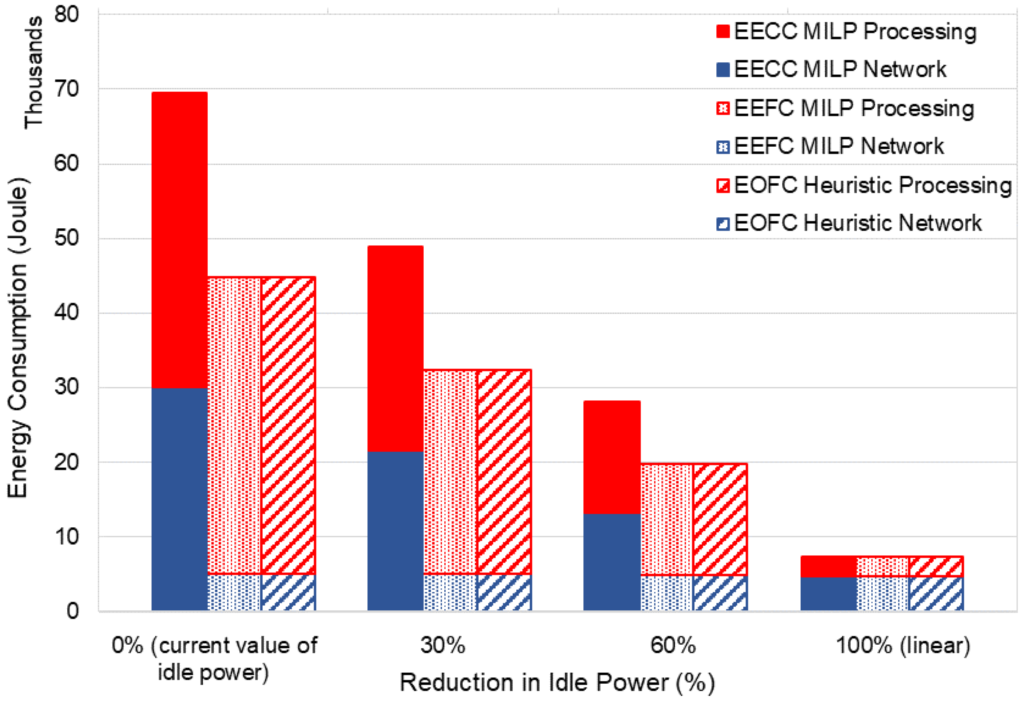


Figure 4.9: Energy consumption of networking equipment and processing in the GPON architecture with varying idle power percentages

Table 4.4: Energy-saving and Energy-increase in the EEFC model compared to the EECC model, with varied percentages of idle power

Energy Type	Percentage of Idle Power			
	90%	60%	30%	0%
Network Saving	83.1%	77.1%	63.5%	0.3%
Processing Increase	0.53%	0.52%	0.47%	0%

Figure 4.9 illustrates the energy consumption of networking equipment and processing for the EECC model, EEFC model and EOFC heuristic in the GPON network with different percentages of idle power. Figure 4.9 also shows that the energy consumption of networking equipment and processing in the EOFC heuristic are the same as in the EEFC model. The energy consumption of networking equipment and processing for both the EEFC model and the EECC model are decreasing with the decreasing percentage of idle power, as shown in Figure 4.9. This is because the idle power dominates the energy consumption of networking equipment and processing server compared to its proportional load power as the size of data used in this work is small.

Figure 4.9 also shows that the energy consumption of networking equipment in the EEFC model is lower than in the EECC model. This is due to the same two factors, as discussed previously in Section 4.3, which are the size of data traversing the networking equipment at the metro and core layers and the utilisation time of that equipment. However, the energy consumption of processing in the EEFC model is higher than in the EECC model when the idle power is 90%, 60% and 30% as illustrated in Figure 4.9. This is due to the high utilisation time of the processing server to transmit the analysed health data to the clinics and the cloud storage in the EEFC model compared to the

EECC model as discussed in Section 4.3. Meanwhile, the energy consumption of processing in the EEFC model and the EECC model are the same when the idle power is 0%. This is because the processing server in the EEFC model and the EECC model served the same number of patients with the same processing and analysis time.

Table 4.4 summarised the energy saving of networking equipment and the energy increase for processing in the EEFC model when compared to the EECC model for all percentages of idle power. The results also show that the energy saving of networking equipment obtained in the EEFC model compared to the EECC model decrease with the decreasing percentage of idle power. This is because, in the EECC model, the total utilisation time of the networking equipment that includes both the time to send the raw ECG signal for processing and analysis and the time to send the analysed health data for feedback and storage purposes are higher than in the EEFC model. Due to this, the energy consumption of networking equipment in the EECC model has a higher energy reduction than the EEFC model with a low percentage of idle power. This reduced the energy saving of networking equipment in the EEFC model compared to the EECC model. Meanwhile, Table 4.4 also shows that the energy increase for processing in the EEFC model compared to the EECC model decrease with the decreasing percentage of idle power. This is because the utilisation time of the processing server (i.e. the time to send the analysed health data for feedback and storage) in the EEFC model is higher than the EECC model. Due to this, the reduction of energy consumption of processing server in the EEFC model with a low percentage of idle power is higher than in the EECC model. Note that the decrease of idle power only affected the energy consumed due to receiving the raw ECG signal from patients, to transmit the analysed data for feedback and permanent storage purposes as shown in Equation (3-28).

4.6 Impact of increasing traffic on EEFC

In this section, the impact of increasing the traffic in the network on the energy consumption of networking equipment and processing in the EEFC model is evaluated by increasing the number of patients from 10% to 90% of the total number of patients for each clinic in 2014/2015 in 10% increments. Note that increasing number of patients increases the traffic in the network. We maintain the processing and analysis time of each processing server at 6.02s where the maximum number of patients each processing server can serve is 669. Therefore, increasing the number of patients in the network will increase the number of utilised processing servers. We consider two scenarios related to the number of processing servers that can be served at each candidate node (i.e. fog node). In the first scenario, each candidate node can serve only one processing server, hence the ϕ_d is parameter i.e. $\phi_d = 1$. The first scenario is applicable for the EEFC model only. In the second scenario, each candidate node can serve more than one processing server, hence ϕ_d is a variable. The second scenario is applicable for both the EECC model and the EEFC model. Also, as we allow each candidate node to serve more than one processing server, therefore, there is additional networking equipment which are Ethernet switches dedicated for healthcare applications to connect the processing servers to each candidate node as shown in Figure 4.10. Note that for scenario 1, we utilised the same network architecture as shown in Figure 3.2

(i.e. without Ethernet switch). Table 4.5 shows the power consumption of the considered Ethernet Switch for scenario 2.

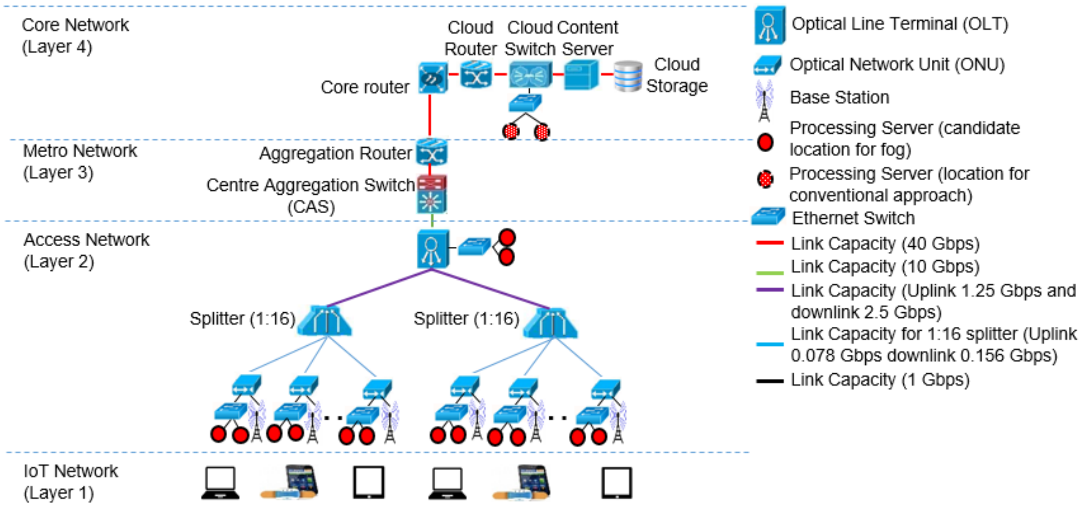

Core Network
(Layer 4)
Cloud Router
Cloud Switch
Content Server
Cloud Storage
Core router
Metro Network
(Layer 3)
Aggregation Router
Centre Aggregation Switch (CAS)
Access Network
(Layer 2)
Splitter (1:16)
Splitter (1:16)
IoT Network
(Layer 1)
Optical Line Terminal (OLT)
Optical Network Unit (ONU)
Base Station
Processing Server (candidate location for fog)
Processing Server (location for conventional approach)
Ethernet Switch
Link Capacity (40 Gbps)
Link Capacity (10 Gbps)
Link Capacity (Uplink 1.25 Gbps and downlink 2.5 Gbps)
Link Capacity for 1:16 splitter (Uplink 0.078 Gbps downlink 0.156 Gbps)
Link Capacity (1 Gbps)

Figure 4.10: GPON architecture with fog computing and Ethernet switches for scenario 2

Table 4.5: Power consumption and capacity of Ethernet switch

Type of device	Maximum Power (Watts)	Idle Power (Watts)	Capacity (Gbps)
Ethernet switch	3.52	0.57	16

The same MILP model in Section 3.5 for FOA and Section 3.6 for CA are utilised to evaluate the performance of both approaches under GPON network. Table 4.6 shows the additional parameters and variables included in the model. For scenario 2, as Ethernet switches are used to connect more than one processing server to the ONU and OLT, Equation (4-2) replace Equation (3-25) and Equation (4-3) replace Equation (3-75) for FOA and CA to include the energy consumed by the Ethernet switches, respectively. Meanwhile, additional Equations (4-4)-(4-7) are used to calculate the energy consumed by the Ethernet switch to serve the raw health data traffic, analysed health data feedback traffic and analysed health data storage traffic, respectively. Note that the energy of the Ethernet switches is consumed if the utilised processing servers are connected to it.

Table 4.6: Additional parameters for Ethernet switch

Parameter	
PES	Maximum power consumption of an Ethernet switch (W)
IES	Idle power consumption of an Ethernet switch (W)
CES	Maximum capacity of an Ethernet switch (bps)
Variables	
$ETES$	Total energy consumption of Ethernet switches
$EESP$	Energy consumption of Ethernet switches required to relay raw health data traffic
$EESF$	Energy consumption of Ethernet switches required to relay analysed health data feedback traffic
$EESS$	Energy consumption of Ethernet switches required to relay analysed health data storage traffic

$$EFN = EPS\ c + ETES\ \eta \tag{4-2}$$

$$ECSN = (EPS + ETES)\ c \tag{4-3}$$

$$ETES = EESP + EESF + EESS \tag{4-4}$$

$$EESP = \sum_{i \in FN} \left(IES \; x \; Y_i + P_i \; \frac{(PES - IES)}{CES} \right) \tau a \tag{4-5}$$

$$EESF = \sum_{i \in FN} \left(IES \; x \; Y_i + F_i \; \frac{(PES - IES)}{CES} \right) \tau b \tag{4-6}$$

$$EESS = \sum_{i \in FN} \left(IES \; x \; Y_i + S_i \; \frac{(PES - IES)}{CES} \right) \tau c \tag{4-7}$$

In addition, similar input parameters in Section 4.2.4 under the GPON network have been employed for scenario 1 to evaluate the performance of the EEFC model in terms of energy consumption of networking equipment and processing under increasing traffic. Meanwhile for scenario 2, similar input parameters in Section 4.2.4 under GPON network, except for the data rate for permanent storage (δc) and its transmission time (τc), are employed for the EECC model and EEFC model. This is because, in scenario 2, the data rate per patient to send the analysed data to the cloud for permanent storage for the EECC model (i.e. CA) and EEFC model (i.e. FOA) decreases with increasing number of patients, which, in turn, increases its transmission time. Increasing the number of patients in the network also reduced the data rate for feedback (δb) and increases its transmission time (τb) for the EECC model (i.e. scenario 2) to 336 bps and 0.76 ms, respectively. The values remain the same for all increasing percentage of patients as the allocated data rate is the minimum rate in the LTE when using the QPSK modulation scheme. Table 4.7 shows the data rate for permanent storage, $\delta c,$ and its transmission time, $\tau c,$ for the EEFC model (i.e. FOA) and EECC model (i.e. CA) for scenario 2 with the increasing number of patients in the network.

Table 4.7: Data rate and transmission time for permanent storage with varying numbers of patient in the network for the EECC and EEFC model under scenario 2

Approach	CA		FOA	
	δc (kbps)	τc (s)	δc (kbps)	τc (s)
10%	7.317	0.035	0.317	0.81
20%	6.708	0.038	0.291	0.88
30%	6.199	0.041	0.269	0.95
40%	5.775	0.044	0.250	1.02
50%	5.346	0.048	0.232	1.1
60%	5.037	0.051	0.218	1.17
70%	4.741	0.054	0.205	1.25
80%	4.492	0.057	0.194	1.31
90%	4.245	0.060	0.184	1.39

Due to the complexity of evaluating the MILP model for a network of large size for each increasing percentage of traffic, the EOFC heuristic is used to study the performance of the energy consumption of networking equipment and processing for the EEFC model compared to the EECC model.

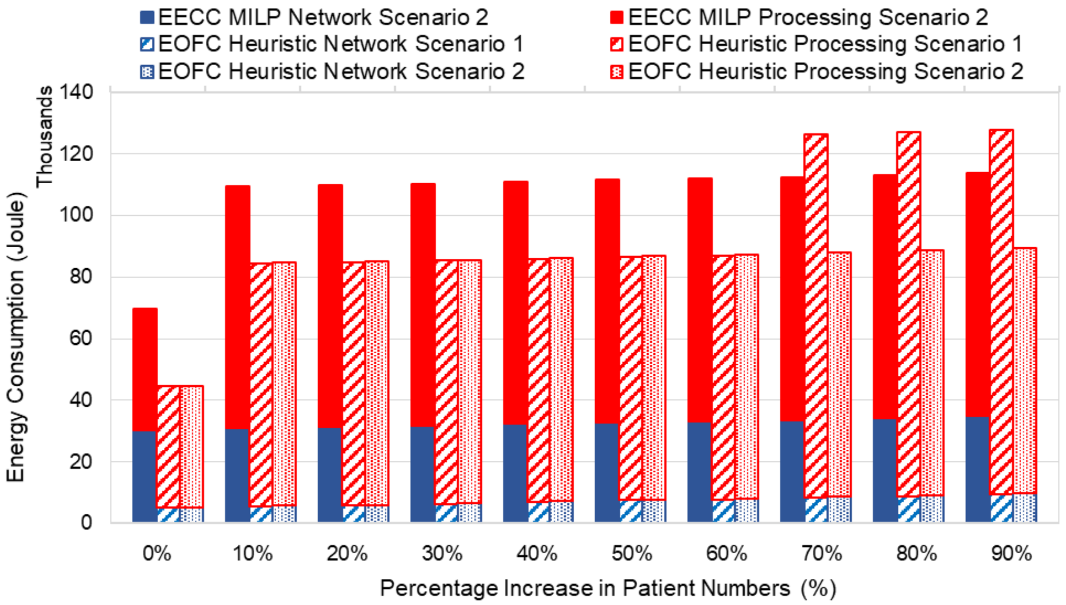


Figure 4.11: Energy consumption of networking equipment and processing in EOFC heuristic under scenario 1 and EECC model and EOFC heuristic under scenario 2 with the increasing number of patients

Table 4.8: Energy-Saving and Energy-Increased in EOFC heuristic under scenario 1 and scenario 2 compared to the EECC model under scenario 2, for varied percentages of increasing traffic.

Percentage of Increasing Traffic (%)	Scenario 1		Scenario 2	
	Network Saving (%)	Processing Increase (%)	Network Saving (%)	Processing Increase (%)
0	83.1	0.53	83.1	0.53
10	81.7	0.34	81.1	0.38
20	81.3	0.34	80.6	0.42
30	79.8	0.34	79.1	0.45
40	78.4	0.34	77.7	0.48
50	77.0	0.34	76.3	0.52
60	76.6	0.33	75.9	0.55
70	75.3	33.27	74.5	0.59
80	74.1	33.25	73.2	0.62
90	72.8	33.23	71.9	0.65

Figure 4.11 shows the energy consumption of networking equipment and processing for EOFC heuristic for scenario 1 and EECC model and EOFC heuristic for scenario 2, when the traffic is based on 2014/2015 (i.e. 0% increase) and increased by 10% to 90% from the total number of patients for each clinic in 2014/2015 in 10% step units. Table 4.8 shows the percentage networking equipment energy saving and energy increase in processing in EEOF heuristic of both scenarios compared to the EECC model of scenario 2 when the increase in percentage of patients is 0% to 90%.

The results in Figure 4.11 show that the total energy consumption of the EECC model (i.e. scenario 2) and EOFC heuristics (i.e. scenario 1 and scenario 2) increase with an increasing percentage of patients. The increase in energy is a result of increasing the number of patients which, in turn, has increased both the total traffic traversing the network and the total number of utilised networking and processing equipment. The results in Figure 4.11 show that the total energy consumption of both networking equipment and processing in the EOFC heuristic of scenario 1 is lower than the EECC model when the percentage increase in patients is equal to or less than 60%. Meanwhile, for the EOFC heuristic under scenario 2, the total energy is lower than the EECC model for all percentages of increasing traffic. The low total energy consumed in the EOFC heuristic of both scenarios is mainly due to the low energy consumed by the networking equipment as a result of the small number of utilised networking equipment and its utilisation time in the EOFC heuristics compared to the EECC model as explained in Section 4.3.

The results also show that the total energy consumed by both networking equipment and processing in the EOFC heuristic of scenario 1 is higher than the EECC model when the percentage of traffic increase is more than 60%. This is because of the increase in the number of processing servers utilised in the EOFC heuristic of scenario 1 due to the limited link capacity at the access network, hence increasing the energy consumption of processing of the EOFC heuristic. Note that, the locations of the processing servers are at both OLT and ONUs when the percentage increase in patients is more than 60%. The percentage energy increase of processing in the EOFC heuristic compared to the EECC model are as shown in Table 4.8.

Figure 4.11 also shows that, the energy consumption of networking equipment in the EOFC heuristic under scenario 2 is slightly higher than in scenario 1. This is due to the additional energy consumed by the Ethernet switches at the access layer and the increasing utilisation time of the networking equipment to send the analysed health data storage traffic to the cloud storage in scenario 2, hence high energy is obtained for EOFC heuristic of scenario 2.

Table 4.8 also shows that the energy saving of networking equipment in the EOFC heuristic of both scenarios compared to the EECC model decreases with increasing percentage of traffic. This is because the increase in energy consumption of networking equipment in EOFC heuristic of both scenarios is higher than the EECC model. For instance, in scenario 1, more networking equipment is utilised to place the processing servers (OLT and ONU) compared to the EECC model. Meanwhile, for scenario 2, the utilisation time of the networking equipment to perform storage tasks in the EOFC heuristic is higher than in the EECC model. Besides, this is also due to the increasing energy consumption of the BS and ONU in both approaches to serve the raw health data traffic (EOFC heuristic of scenario 1 and EECC model and EOFC heuristic of scenario 2) and, hence, reduces the energy saving of networking equipment in the EOFC heuristic when compared to the EECC model.

Figure 4.11 also shows that the energy consumption for processing in the EOFC heuristic in scenario 1 and scenario 2 is slightly higher than in the EECC model for all percentages of increasing traffic. Table 4.8 shows that the energy increase for processing in EOFC heuristic of scenario 1 and scenario 2 compared to the EECC model with 10% of increasing traffic reduced to 0.34% and 0.38%, respectively, compared to traffic in 2014/2015 which is 0.53% (i.e. 0% of increasing traffic). This is due to the increasing utilisation time of the processing server in the EECC model to perform the feedback task since the increasing number of patients reduced the data rate allocated to each patient for the feedback transmission. However, the percentage energy increase for processing in the EOFC heuristic compared to the EECC model under scenario 2 increases with increasing traffic from 10% to 90%, as shown in Table 4.8. This is because the increasing total utilisation time of the processing servers in the EOFC heuristic under scenario 2 is higher than in the EECC model for all percentages of increasing traffic. Therefore, this increases the percentage energy increase in EOFC heuristic compared to the EECC model under scenario 2. Meanwhile, Table 4.8 also shows that the energy increase in processing for EOFC heuristic under scenario 1 is the same when the percentage increase in patients ranges from 10% to 50%. This is mainly due to the same utilisation time of the processing servers to receive the raw health data and to transmit the analysed health data for feedback and storage in the EOFC heuristic of scenario 1 and the EECC model, regardless of the increasing number of patients.

4.7 Impact of different access networks on EEFC model

In this section, the performance of the EEFC model (FOA), compared to the EECC model (CA), under the Ethernet network is evaluated in terms of energy consumption of both networking equipment and processing. The fog architecture and the parameters considered for health monitoring applications using the Ethernet access network are also provided.

4.7.1 Fog computing architecture for health monitoring applications under Ethernet access network

The Ethernet architecture in the fog network is shown in Figure 4.12. The only difference between the Ethernet and GPON network is at the access layer where in the Ethernet network, the ONU and OLT are replaced with an access switch (ASW) and an aggregation switch (AGS), respectively. The AGS in the Ethernet network has 62 ports [135]. To maintain resilience, each ASW will be connected to 2 ports at the AGS, hence only 31 ASWs can be connected to the AGS.

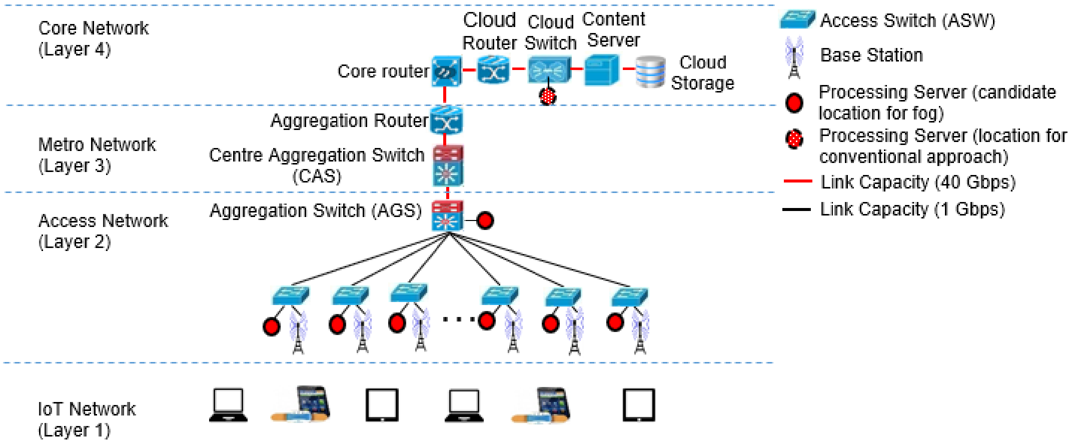


Figure 4.12: Ethernet architecture in the fog network

The network connection at the access layer in the Ethernet network is optimised using MALTAB, as in Section 4.2.1, and the optimal location of AGS for the Ethernet network is the same as the optimal location of the OLT in Figure 4.2, as represented by a black diamond. Note that the optimised location of AGS had the lowest total distance from 26 ASWs.

4.7.2 MILP model and parameters consideration under Ethernet access network for health monitoring system

In order to optimise the energy consumption of networking equipment and processing under the Ethernet Network, an additional set, parameters and variables to those in Table 3.1 and Table 3.2 related to the Ethernet specification are summarised in Table 4.9 (also can be found in Appendix 1). Note we consider that each candidate node can serve only one processing server, hence the ϕ_d parameter value, i.e. $\phi_d = 1$.

Table 4.9: Set and parameters for Ethernet network

Set	
ASW	Set of access switches
Parameters	
$PASW$	Maximum power consumption of an access switch (W)
$IASW$	Idle power consumption of an access switch (W)
$CASW$	Maximum capacity of an access switch (bps)
$PAGS$	Maximum power consumption of an aggregation switch (W)
$IAGS$	Idle power consumption of an aggregation switch (W)
$CAGR$	Maximum capacity of an aggregation switch (bps)
Variables	
$ETASW$	Total energy consumption of access switches
$EASWP$	Energy consumption of access switches required to relay raw health data traffic
$EASWF$	Energy consumption of access switches required to relay analysed health data feedback traffic
$EASWS$	Energy consumption of access switches required to relay analysed health data storage traffic
$ETASG$	Total energy consumption of aggregation switches
$EASGP$	Energy consumption of aggregations switches required to relay raw health data traffic
$EAGSF$	Energy consumption of aggregation switches required to relay analysed health data feedback traffic
$EAGSS$	Energy consumption of aggregation switches required to relay analysed health data storage traffic

We use the MILP model in Section 3.5 with several modifications at the access network due to the different types of networking devices at the access layer between the GPON and the Ethernet access network. The energy consumption of access network in Equation (3-3) is redefined as follows:

$$EAN = (ETBS + ETASW + ETAGS)\ \eta \tag{4-8}$$

where the total energy consumption of base stations, $ETBS,$ together with the energy consumed by the BS to serve the raw health data traffic and analysed health data feedback traffic are the same as in Equation (3-4)-(3-6). Meanwhile, Equation (3-7)-(3-10) for ONUs are replaced with Equation (4-9)-(4-12) to consider access switches, respectively, as given below:

$$ETASW = EASWP + EASWF + EASWS \tag{4-9}$$

$$EASWP = \sum_{i \in ASW} \left(IASW\ x\ \zeta a_i + P_i\ \frac{(PASW - IASW)}{CASW} \right) \tau a \tag{4-10}$$

$$EASWF = \sum_{i \in ASW} \left(IASW\ x\ \zeta b_i + F_i\ \frac{(PASW - IASW)}{CASW} \right) \tau b \tag{4-11}$$

$$EASWS = \sum_{i \in ASW} \left(IASW\ x\ \zeta c_i + S_i\ \frac{(PASW - IASW)}{CASW} \right) \tau c \tag{4-12}$$

Also, Equation (3-11)-(3-14) for OLTs are replaced with Equation (4-13)-(4-16) to consider aggregation switches, respectively.

$$ETAGS = EAGSP + EAGSF + EAGSS \tag{4-13}$$

$$EAGSP = \sum_{i \in AGS} \left(IAGS\ x\ \zeta a_i + P_i\ \frac{(PAGS - IAGS)}{CAGS} \right) \tau a \tag{4-14}$$

$$EAGSF = \sum_{i \in AGS} \left(IAGS\ x\ \zeta b_i + F_i\ \frac{(PAGS - IAGS)}{CAGS} \right) \tau b \tag{4-15}$$

$$EAGSS = \sum_{i \in AGS} \left(IAGS\ x\ \zeta c_i + S_i\ \frac{(PAGS - IAGS)}{CAGS} \right) \tau c \tag{4-16}$$

To determine the data rate and transmission time to transmit the ECG signal to the processing server for processing and analysis, and to transmit the analysed ECG data to the clinic for feedback, as well as to transmit the analysed ECG data to cloud for storage under Ethernet network, we used the same methodologies as explained in Section 4.2.4. Note that under the Ethernet network, we considered the link between the ASW and CAS to be the minimum shared link capacity to determine the data rate for feedback (i.e. FOA and CA) and storage (i.e. FOA). This is due to the same reason as explained in Section 4.2.4. Table 4.10 shows the related parameter inputs calculated for FOA and CA under the Ethernet network. Meanwhile, Table 4.11 summarised the additional input parameters related to the Ethernet specification to those in Table 4.3.

Table 4.10: Parameter inputs for FOA and CA

Parameter	FOA	CA
Transmission time to transmit raw ECG data to processing server, τa (s)	188.1	188.1
Data rate to transmit raw ECG data to processing server, δa (kbps)	1.344	1.344
Transmission time to transmit analysed ECG data to clinic, τb (s)	0.059	0.059
Data rate to transmit analysed ECG data to clinic, δb (kbps)	4.368	4.368
Transmission time to transmit analysed ECG data to cloud storage, τc (s)	0.057	0.032
Data rate to transmit analysed ECG data to cloud storage, δc (kbps)	4.484	8.071

Table 4.11: Input parameters for the Ethernet network

Parameter	Value
Maximum power consumption of access switch, $PASW$	40 W [136]
Access switch capacity, $CASW$	20 Gbps [136]
Maximum power consumption of aggregation router, $PAGS$	728 W [137]
Aggregation switch capacity, $CAGS$	120Gbps [135]

4.7.3 Results and analysis of EEFC model under Ethernet access network for health monitoring system

In this section, the performance of the EECC model, EEFC model and EOFC heuristic are evaluated in term of energy consumption of networking equipment and processing under the Ethernet network.

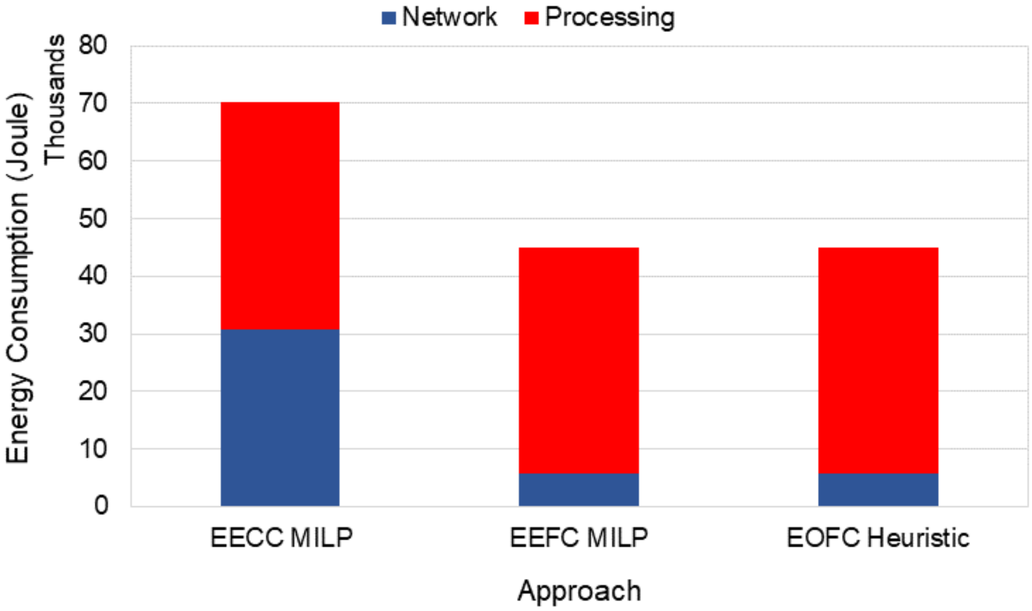


Figure 4.13: Energy consumption of networking equipment and processing in Ethernet architecture

Figure 4.13 illustrates the energy consumption of networking equipment and processing under the Ethernet network for the EECC model, EEFC model and EOFC heuristic. We used the EECC model as our benchmark to evaluate the performance of energy consumption in both the EEFC model and EOFC heuristic. The energy saving of networking equipment in the EEFC model, compared to the EECC model under Ethernet network, is 81.5% as illustrated in Figure 4.13 where the energy consumption of networking equipment in the EEFC model is lower than the EECC model. This saving is due to the same two factors, as explained in Section 4.3. The figure also shows that the EOFC heuristic has the same performance as the EEFC model in terms of network energy consumption. This is due to the fact that the same location, which is AGS, is used to place the processing server in both models. Therefore, the EOFC heuristic has the same energy saving as in the EEFC model compared to the EECC model.

However, the energy saving of networking equipment obtained in the EEFC model compared to EECC model under the Ethernet network is lower than in the GPON network (i.e. 83.1%). Also, the increase in the energy consumption of networking equipment for the EECC model and the EEFC model using the Ethernet network compared to the GPON network is 2.3% and 11.8%, respectively. This is mainly due to the low power consumption of the total ONU and OLT in the GPON network which is 96.4% less when compared to the total power consumption of ASW and AGS in the Ethernet network under full utilisation (maximum power). Note that the power consumption of ONU is 80% lower than the ASW while OLT is 97.3% lower than the AGS. It is worth noting that the utilisation time of the networking equipment to send the analysed health data feedback traffic and analysed health data storage traffic in the GPON network is high when compared to the Ethernet network due to the link capacity constraint between the ONU and the OLT in the GPON network. The high utilisation time increases the energy consumption of the networking equipment to perform the feedback and storage tasks in the GPON network. However, the energy efficiency of the ONU and OLT dominates the increasing energy due to the high utilisation time to perform those tasks. Hence there is a high energy saving of networking equipment under the GPON network compared to the Ethernet network.

Figure 4.13 also shows that the energy consumption of processing in the EEFC model and EOFC heuristic is slightly higher than the EECC model by 0.01%, mainly due to the same reason as explained in Section 4.3 for the GPON network. However, the increase in energy processing in the Ethernet network is low when compared to the GPON network, mainly due to the low utilisation time of the processing server to send the analysed data to the clinic and cloud storage in the EEFC model and EOFC heuristic, resulting from the high link capacity under Ethernet network compared to the GPON network.

4.8 Conclusions

This chapter has investigated the impact of integrating fog computing at the network edge on the energy consumption of networking equipment and processing for health monitoring applications. This is accomplished by deploying a processing server at the network edge to perform both processing and analysis of health data. Realistic locations for base station, OLT, clinic as well as real number of patients are considered. A MILP model (EEFC) and a heuristic (EOFC) to optimise the location of the processing server at the access layer were developed. The result of the EEFC model and the EOFC heuristic reveal that the optimal location for placing the processing server is at the OLT as it is the nearest shared point to the patients. As a result, there is 83.1% of network energy-saving in the EEFC model when compared to the EECC model where the processing is performed at the central cloud. Hence, this study has shown that by reducing the traffic and the utilisation time of the networking equipment by using fog processing at the network edge the energy consumption of the transport network can effectively be reduced. Nonetheless, 0.53% of processing energy-increase was observed in the EEFC model in comparison to the EECC model. However, the total energy consumption of networking equipment and processing in the EEFC model is 35.7% less when compared to the EECC model. Note that the energy saving of networking equipment and the energy increase of processing in EOFC heuristic are equal to those of the EEFC model. We also studied the impacts of decreasing idle power of devices and increasing volume of traffic in the network upon the performance of the EEFC model and EOFC heuristic. The results revealed that the performance of the proposed EEFC model and EOFC heuristic are the same and are more energy efficient when compared to the EECC model. However, the energy efficiency of the EEFC model is limited by the link capacity constraints at the access network where more processing servers are required to serve the increasing patients when the number of processing servers at each fog node is limited. Also, the results show that the integration of fog computing with an Ethernet network also exhibited 81.3% network energy-saving, in comparison to the conventional approach.

Chapter 5

Energy-efficient fog computing with Long Term Evolution for machine (LTE-M) for fall monitoring applications

5.1 Introduction

In this chapter, the use of fog computing for fall monitoring applications is investigated considering a realistic sample of elderly patients at West Leeds, United Kingdom, who suffer from heart diseases, to be the respondents. The respondents among the elderly are selected because falls are the leading cause of injuries and deaths among seniors [138], [139]. It has been reported that one-third of the elderly people aged 65 years and above fall each year [140]. Meanwhile, the REGARDS study found that the presence of a history of both heart disease (i.e. atrial fibrillation) and falls is associated with a significantly higher risk of mortality [141]. Therefore, immediate treatment is necessary to save the patients.

Among all the available medical monitoring services, video analysis happens to be the most common approach to detect a fall [142]. It has been reported that falls account for 10% – 25% of the ambulance call-outs for elderly people, which cost £115 per call-out [140]. Therefore, to avoid a false alarm given to the doctors, two stages of detection are performed as follows: At an event of a patient fall, the accompanying IoT device will first detect it on the basis of its limited video processing capabilities. Then, it will send a 15-s video recording as proposed in [7] to the fog servers to reconfirm the occurrence of a patient fall on the basis of considerably higher processing capabilities than the IoT devices before triggering a doctor call. Thus, the 15-

s video recording will be processed and analysed at the processing servers in the fog layer to reconfirm the occurrence of a fall.

Because of the complexity of the model, an extreme scenario is considered wherein the events trigger sending a video for each patient at the same time. In this work, the placement of the processing servers in the network is optimised by using the same mixed integer linear programming (MILP) model (Section 4.7) and the heuristic model (Section 4.4) discussed in Chapter 4, so that the total energy consumption of fall monitoring is minimised.

5.2 Parameters consideration

This section elaborates the methodologies of determining the model input parameters considered for fall monitoring applications. The input parameters are divided into several types such as the number of monitored patients in West Leeds, UK, the processing and analysis time of health data (i.e. video recording signals), and the calculation of the data rate and the traffic transmission.

5.2.1 Total number of monitored patients in West Leeds, UK

In this study, 37 clinics located at West Leeds were considered to monitor the elderly patients with heart disease, similar to Chapter 4. The total number of patients of all ages who suffered from heart disease was determined on the basis of the data from Public Health England [143]. As reported in [140], the percentage of elderly people aged 65 years and above was 17.7% and one-third of them experienced falls each year. Therefore, 17.7% of the total patients with heart disease in each clinic were considered elderly people who suffered from heart disease, and only a third of them were monitored, which reflected the traffic demands in this study. Table 5.1 presents the deduced total number of elderly patients registered at each clinic who were expected to experience a fall.

Table 5.1: Number of monitored elderly patients in clinics expected to experience a fall

Clinic	Number of Patients	Clinic	Number of Patients
Craven Road Medical Practice	3	Leeds Student Practice	0
Hyde Park Surgery	1	Burton Croft Surgery	4
Laurel Bank Surgery	1	Kirkstall Lane Medical Centre	1
Burley Park Medical Centre	4	Thornton Medical Centre	5
Gildersome Health Centre	2	The Dekeyser Group Practice	8
Leigh View Medical Practice	6	West Lodge Surgery	13
Hillfoot Surgery	2	Dr KW McGechaen & Partner	2
Pudsey Health Centre	4	Robin Lane Medical Centre	6
Dr S M Chen & Partner	2	Beech Tree Medical Centre	1
Hawthorn Surgery	3	Priory View Medical Centre	6
High Field Surgery	3	Abbey Grange Medical Centre	4
Vesper Road Surgery	2	Fieldhead Surgery	1
Manor Park Surgery	7	The Highfield Medical Centre	2
Dr G Leeds & Partners	4	Dr F Gupta’s Practice	1
Guiseley and Yeadon Medical Practice	6	Park Road & Menston	6
Yeadon Tarn Medical Practice	4	Rawdon Surgery	4
Dr KJ Manock & Partners	11	Whitehall Surgery	2
Dr JA Browne’s Practice	6	Dr N Saddiq’s Practice	1
Dr JJ McPeakes Practice	2		

5.2.2 Time measurement for video data processing

In this study, each patient transmitted his/her 15-s video recording having a size of 3.36 Mbits to the network by using Kinect's IR sensor with a 640 × 480 resolution at 30 frames per second, as proposed in [144]. The time to process and analyse the video data with a 2.4-GHz processor was around 0.3 ms – 0.4 ms per frame [144]. In this work, 0.4 ms was used as the per frame processing time. Therefore, the duration to process and analyse one video recording per patient was 0.18 s, as calculated below:

$$\tau ps = 15\, s\ \cdot\ 30\, frames/s\ \cdot 0.4\, ms/frame \tag{5-1}$$

5.2.3 Data rate calculation for traffic transmission in the network

As discussed in Chapter 4, we considered 4 min, τt, as the maximum duration to save elderly patients who had heart disease, as proposed by American Heart Association [57] and experienced a fall. This duration, together with the data volume to be transmitted, was used to calculate the data rate needed for each patient. The 4 min included the 15 s of the video recording (i.e. 3.36 Mbits of data) for monitoring τm, the transmission time to send the video recording to the processing server τmax, the transmission time to send the analysed health data feedback traffic to the clinic τb, and the time to perform the processing and the analysis τp.

The time for processing and analysis was calculated on the basis of the number of patients that each processing server could serve (Pat) and the duration to process a video recording per patient (τps), which was 0.18 s, as calculated in Section 5.2.2; thus, $\tau p = \tau ps \cdot Pat$. Two scenarios were considered to evaluate the energy consumption of the networking equipment and processing as follows:

1) Limited number of patients per processing server (scenario 1)

In this approach, five scenarios were considered where the Pat value was 20%, 40%, 60%, 80%, and 100% of the total patients considered in this work. For each Pat, the number of processing servers that could be served at each candidate node was unlimited. Therefore, the total number of patients that could be served at each candidate node $Patm$ was equal to the total number of patients considered in the network, as all the processing servers could be placed at the same node.

2) Limited number of processing servers per candidate node (scenario 2)

In this approach, a single processing server was assigned to serve 20% of the total patients (Pat), and the number of processing servers that could be served at each candidate node N was limited. Five scenarios were investigated, where the N value was 1, 2, 3, 4 and 5, to investigate the distribution of the processing servers in the network with a limited number of processing servers per candidate node. Therefore, the total number of patients that could be served at each candidate node was $Patm = Pat \cdot N$.

In this work, videos were assumed to be processed in series. Therefore, the worst-case scenario was considered to be one in which all the videos were processed and analysed before the feedback was sent. To determine the time required to transmit the analysed data to the clinics for feedback, the processing servers at each candidate node assigned to serve $Patm$ were considered. Then, the minimum shared link $Lmin$ at the edge network provisioned for healthcare applications where the processing server was located (i.e. the uplink and the downlink between the ONU and the OLT for FOA and CA, respectively) was determined. By considering all the patients to be served by the processing servers located at the same candidate node, we divided this minimum link capacity equally to each patient, hence obtaining a data rate of $\delta f = Lmin/Patm$ for each patient. Note that the reason to limit the feedback data rate by the data rate available for healthcare applications in the network edge (i.e. GPON links) was to reduce the energy consumed by the base station, as explained in Section 4.2.4. As described in Chapter 4, an LTE-M base station with the QPSK modulation scheme was considered, which yielded a minimum of 336 bps per physical resource block (PRB). Therefore, the number of PRBs for each patient to send the feedback data was $Rf = \lfloor \delta f/336\ bits \rfloor$, where Rf was the minimum integer value to ensure that the link capacity that was provisioned for healthcare in the network was not exceeded.

The maximum size allowed for a notification payload according to Apple Push Notification Services is 256 bytes (i.e. 2.048 kbits) [145]. In this chapter, the size of the analysed video recording, which was 2.048 kbits (α), was to be sent to the clinics for feedback purposes and to be permanently stored in the cloud storage. Therefore, the data rate to send the feedback data was $\delta b = Rf \cdot 336\, bits$, while the transmission time was $\tau b = \alpha/\delta b$. The remaining transmission time to send the video recording to the processing servers was $\tau max = \tau t - \tau m - \tau b - \tau p$, which yielded a minimum data rate of $\delta min = 3.36\, Mbits/\tau max$ to transmit the video signal to the processing server. However, as the data traversed the LTE base station and the minimum allocation of resources to each user was one PRB, the number of PRBs that could be assigned to each patient to transmit his/her video recording was $Rp = \lceil \delta min/336\ bps \rceil$, where Rp was the maximum integer value to ensure that the given data rate was equal to or higher than the minimum required data rate so that the system could work within 4 min. Hence, the data rate to send a video signal to the processing server was $\delta a = Rp \cdot 336\, bps$ with a maximum transmission time of $\tau a = 3.36\ Mbits/\delta a$ per patient.

The data rate to send the analysed data at each processing server to the cloud storage for permanent storage was determined by dividing the lowest shared uplink or node capacity provisioned by a health M2M application from the processing server to the cloud storage, $Cmin$ by $Patm$ (i.e. $\delta c = Cmin/Patm$). Hence, the time required to transmit the analysed video data to the cloud storage was $\tau c = \alpha/\delta a$. As discussed in Chapter 4, 0.3% of the maximum available shared network device and link capacities were considered to be dedicated to our healthcare applications. Table 5.2 shows the input parameter calculated for the FOA and CA, as discussed above for scenario 1, while Table 5.3 shows the input parameter calculated for the FOA for scenario 2. As shown in Table 5.2, the data rate to transmit the video signal to the processing server for FOA was higher than that for CA when the percentage of patients that could be served at each processing server was 80%, while the other parameters were the same. This was because the actual data rate to transmit the video signal to the processing servers was determined on the basis of the number of PRBs (i.e. each with 336 bps) while ensuring that the total data rate given by the total number of PRBs per patient was equal to or higher than the minimum data rate required for each approach so that the system could work within 4 min. Therefore, the same number of PRBs could be given to FOA and CA, although their required minimum data rate was different. However, to reduce the number of utilised base stations, the minimum number of PRBs would be allocated to each patient. Hence, different data rates could also be assigned for each approach.

Table 5.2: Parameter inputs for FOA and CA for scenario 1

Type of Data	Approach	Percentage of Patients per Processing Server				
		20%	40%	60%	80%	100%
Data rate to transmit video to processing server, δa (kbps)	FOA	15.456	15.792	16.128	16.800	17.136
	CA	15.456	15.792	16.128	16.464	17.136
Transmission time to transmit video data to processing server, τa (s)	FOA	217.39	212.77	208.33	200	196.08
	CA	217.39	212.77	208.33	204.08	196.08
Data rate to transmit analysed video to clinics, δb (kbps)	FOA	1.344	1.344	1.344	1.344	1.344
	CA	3.024	3.024	3.024	3.024	3.024
Transmission time to transmit analysed video to clinics, τb (s)	FOA	1.524	1.524	1.524	1.524	1.524
	CA	0.68	0.68	0.68	0.68	0.68
Data rate to transmit analysed video to cloud storage, δc (kbps)	FOA	1.674	1.674	1.674	1.674	1.674
	CA	38.571	38.571	38.571	38.571	38.571
Transmission time to transmit analysed video to cloud storage, τc (s)	FOA	1.223	1.223	1.223	1.223	1.223
	CA	0.053	0.053	0.053	0.053	0.053

Table 5.3: Parameter inputs for FOA and CA for scenario 2

Type of Data	Number of Processing Servers per Candidate Node				
	1 PS	2 PSs	3 PSs	4 PSs	5 PSs
Data rate to transmit video to processing server, δa (kbps)	15.456	15.456	15.456	15.456	15.456
Transmission time to transmit video to processing server, τa (s)	217.39	217.39	217.39	217.39	217.39
Data rate to transmit analysed video to clinics, δb (kbps)	8.064	4.032	2.688	2.016	1.344
Transmission time to transmit analysed video to clinics, τb (s)	0.254	0.508	0.762	1.016	1.524
Data rate to transmit analysed video to cloud storage, δc (kbps)	8.370	4.185	2.79	2.092	1.674
Transmission time to transmit analysed video to cloud storage, τc (s)	0.245	0.489	0.734	0.979	1.223

5.3 Mathematical model for Energy-efficient fog computing health monitoring application (EEFC)

In this section, the same objective function and the MILP model used in Sections 3.5 and 4.6 are utilised to optimise the location of the processing servers at the network edge while minimising the energy consumption of the networking equipment and processing. Note that the total time for processing and analysis of the raw video data, τp_d in Equation (3-60), was obtained on the basis of the total number of patients served by the processing server at node d, and the time for processing and analysis given per video where m and $\acute{c}$ were equal to 0.18 s and 0, respectively, in both scenario 1 and scenario 2. Furthermore, note that for both scenario 1 and scenario 2, ϕ_d was a variable denoting that each candidate node could serve more than one processing server; else, ϕ_d was a parameter where $\phi_d = 1$. For scenario 2, an additional constraint, Equation (5-2), was used to ensure that the number of processing servers deployed at each candidate node did not exceed the maximum number of processing servers allowed at each candidate node. The additional parameters (also can be found in Appendix 1) and Equation (5-2) considered in this work are as shown below:

Table 5.4: Additional parameters for EEFC model

N	Maximum number of processing servers per candidate node

$$\phi_i \leq N \quad ; \ \forall i \in FN \tag{5-2}$$

5.4 Results and analysis of EEFC model and the EOFC heuristic

This section presents the results and analysis of the EEFC model (i.e. FOA) for the fall monitoring application with a limited number of patients per processing server (scenario 1) and a limited number of processing servers per candidate node (scenario 2). As in the previous chapters, AMPL software with CPLEX 12.8 solver running on high-performance computing (HPC) cluster with a 12 core CPU and 64 GB RAM was used as the platform for solving the EEFC models. Furthermore, the results of the EOFC heuristic running on a normal PC with 3.2 GHz CPU and 16 GB RAM are provided for real-time implementation of the EEFC model. The same GPON architecture, as that shown in Figure 4.10, was used to evaluate the performance of the EEFC model and the EOFC heuristic in terms of the energy consumption of both the networking equipment and the processing. The processing server used to perform the processing and the analysis of the video recording data was a 2.4-GHz Intel Core-Duo, and the related power consumption was as shown in Table 5.5.

Table 5.5: Power consumption of 2.4-GHz server

Type of Device	Maximum Power (W)	Idle Power (W)
Intel Core Duo (2.4 GHz) [133]	85	10

5.4.1 Limited number of patients per processing server

In this section, the evaluation performance of both the EEFC model and the EOFC heuristic for an increasing percentage of patients served in each processing server is presented. The conventional approach, the EECC model (i.e. CA) in Section 4.6, was used as the benchmark to evaluate the performance of both the EEFC model and the EOFC heuristic for the fall monitoring applications in terms of the energy consumption of both the networking equipment and the processing. Moreover, the optimisation gaps between the EEFC model and the EOFC heuristic were observed and are presented in this section.

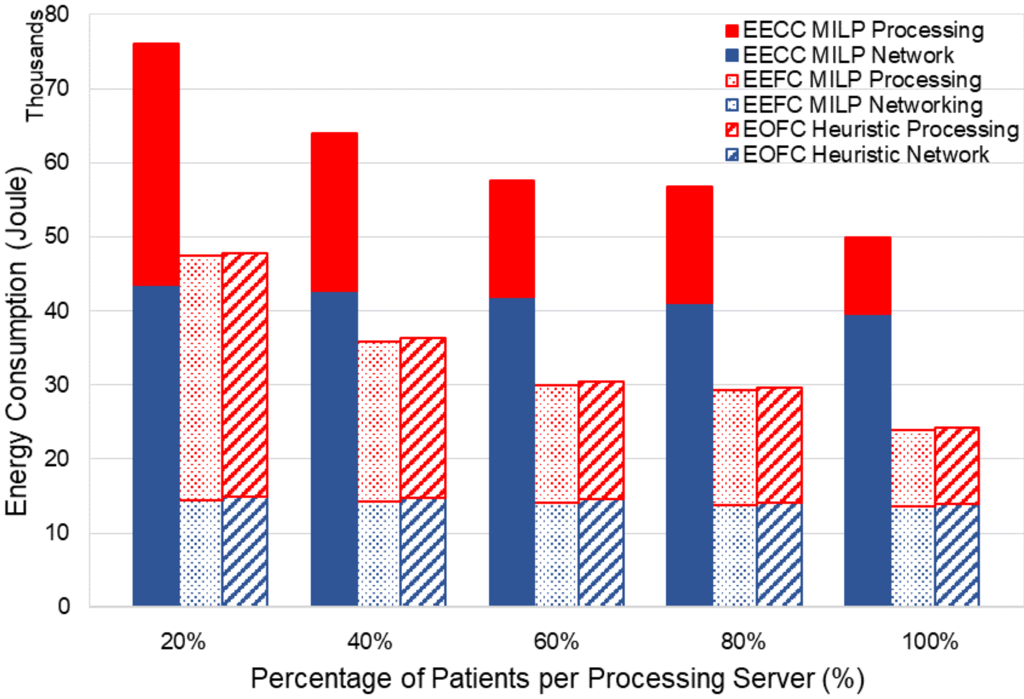


Figure 5.1: Energy consumption of networking equipment and processing for EECC model, EEFC model, and EOFC heuristic for different percentages of patients per processing server

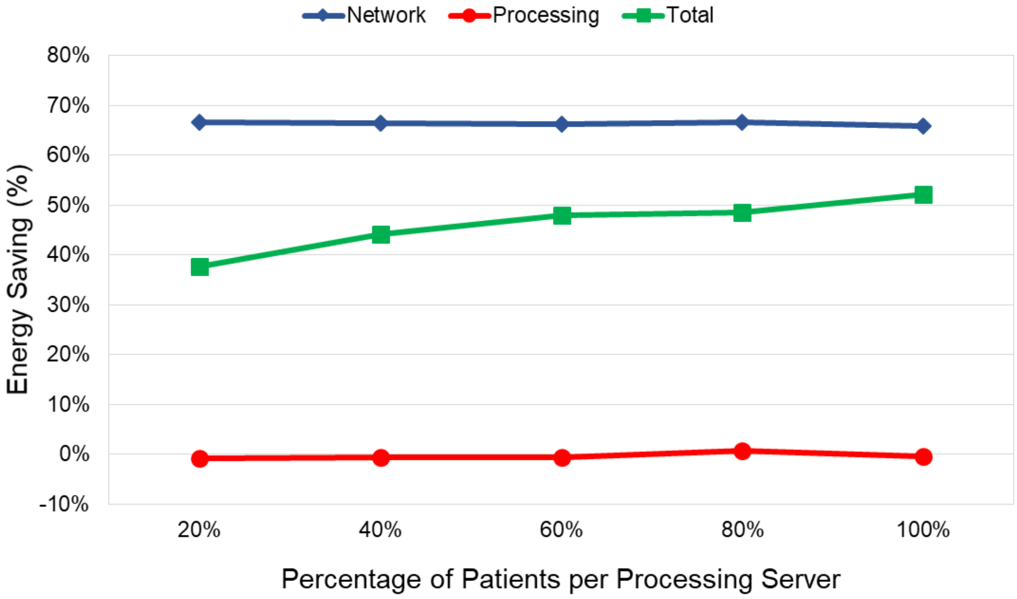


Figure 5.2: Percentage energy saving in EEFC model compared to EECC model for different percentages of patients per processing server

Table 5.6: Optimisation gap between the EEFC model and the EOFC heuristic for different percentages of patients per processing server

	Gap %				
Percentage of patients per processing server	20%	40%	60%	80%	100%
Total energy	0.98%	1.26%	1.47%	1.45%	1.74%
Network energy	3.19%	3.16%	3.14%	3.09%	3.07%
Processing energy	0%	0%	0%	0%	0%

Figure 5.1 shows the energy consumption of the networking equipment and the processing for the EECC model, EEFC model, and the EOFC heuristic, while Figure 5.2 shows the total energy saving, energy saving of the networking equipment, and the energy saving of the processing of the EEFC model as compared to those of the EECC model. The results are shown for an increasing percentage of patients that could be served at each processing server. The results presented in Figure 5.1 revealed that the total energy consumption of the EEFC model was always smaller than that of the EECC model for all percentages of patients per processing server. For instance, the total energy saving of the EEFC model compared with that of the EECC model was 37.7% when a single processing server could serve 20% of the total number of patients in the network, as shown in Figure 5.2. This saving was attributed to the fact that the location of the processing servers in the EEFC model was the OLT, thereby reducing the amount of networking equipment utilised to transmit the raw health data traffic to the processing server. Compared with the EECC model, the location of the processing servers was

in the cloud. Therefore, considerable energy was consumed in the metro and core layers to transmit the raw health data traffic to the processing servers.

Figure 5.2 also shows that when a single processing server could serve 80% of the patients, there was 0.7% processing energy saving in the EEFC model as compared to the EECC model. This conservation was attributed to the low utilisation time of the processing server with the EEFC model to transmit the raw health data traffic to the processing servers as compared to the EECC model. Note that reducing the utilisation time of the processing servers could reduce the energy consumption of the processing. Meanwhile, for the other percentages of patients that could be served by a single processing server, the amount of energy required for the processing in the EEFC model was slightly larger than that in the EECC model, as shown in Figure 5.1. The high energy consumption of processing in the EEFC model was attributed to the high utilisation time of the processing servers to send the analysed health data feedback traffic and the analysed health data storage traffic to the clinics and the cloud storage, respectively, compared to the EECC model, while the same amount of time was used to transmit the raw health data traffic.

Figure 5.1 also shows that the total energy consumption of the EEFC model and the EECC model decreased when more patients could be served by a single processing server. This was because allowing more patients to be served by a single processing server reduced the number of utilised processing servers, thereby reducing the energy consumption required for the processing. Meanwhile, Figure 5.2 shows that the total energy saving increased with an increase in the percentage of patients served by a single processing server. This was because allowing more patients to be served by a single processing server reduced the available time to send the raw video recording to the processing servers, which in turn reduced the amount of energy consumed by the idle power of the networking equipment and the processing server.

Figure 5.2 also shows that the total energy consumption of the EOFC heuristic approached the total energy consumed by the EEFC model. Table 5.6 shows that the overall optimisation gap between the EEFC model and the EOFC heuristic for different percentages of patients per processing server was less than 2%. This gap was mainly attributed to the high energy consumption of the networking equipment in the EOFC heuristic, where the number of utilised base stations with the EOFC heuristic was higher than the EEFC model. Meanwhile, the energy consumption of processing in the EEFC model and the EOFC heuristic was equal because of the same number of utilised processing servers. We also evaluate the computational time needed to run the EOFC heuristic and the EEFC model. The running times of the EOFC heuristic on a normal computer with 3.2 GHz and 16 GB RAM took 15.4 sec to finish which is lower than the running time of the MILP on a high-performance computing (HPC) cluster with a 12 core and 64 GB RAM that took 314 sec to finish when 20% of patients can be served in a single processing server.

5.4.2 Limited number of processing servers per candidate node

In this section, the performance evaluation of the EEFC model and EEFC heuristic for an increasing number of processing servers per candidate node is presented. Moreover, the optimisation gaps between the EEFC model and the EOFC heuristic were observed and are presented in this section. Note that when the number of processing servers per candidate node was limited to 1, the same GPON architecture as that described in Section 3.2, was considered where a single processing server was connected directly to the ONU or the OLT. Meanwhile, for the other number of processing servers per candidate node, the GPON architecture described in Section 4.6 was considered where the Ethernet switch was used to connect the processing servers to the ONU or the OLT.

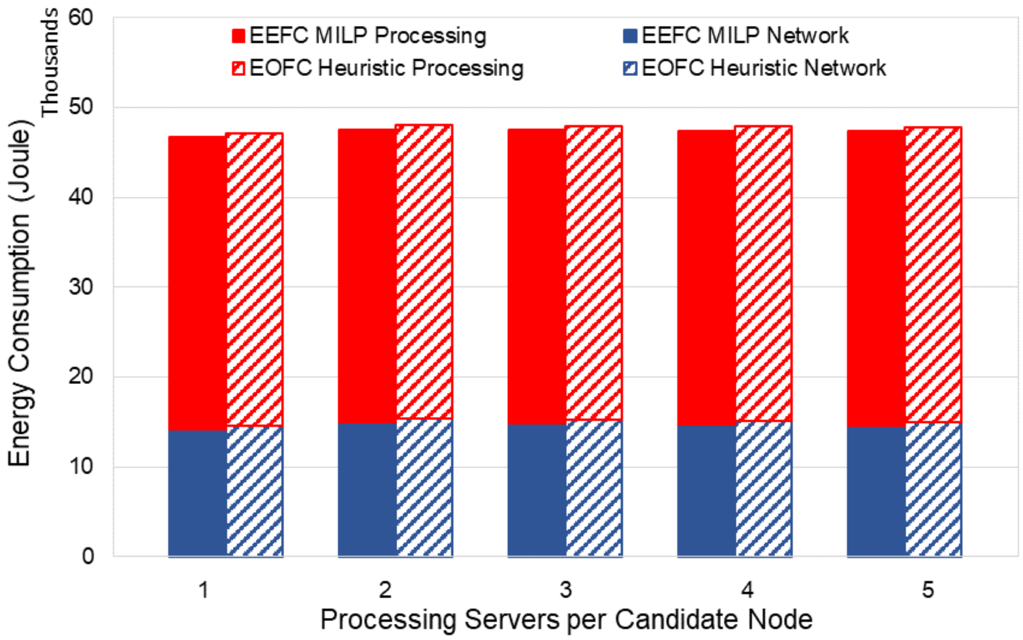


Figure 5.3: Energy consumption of networking equipment and processing for EEFC model and EOFC heuristic for different numbers of processing servers per candidate node when 20% of patients can be served in a single processing server

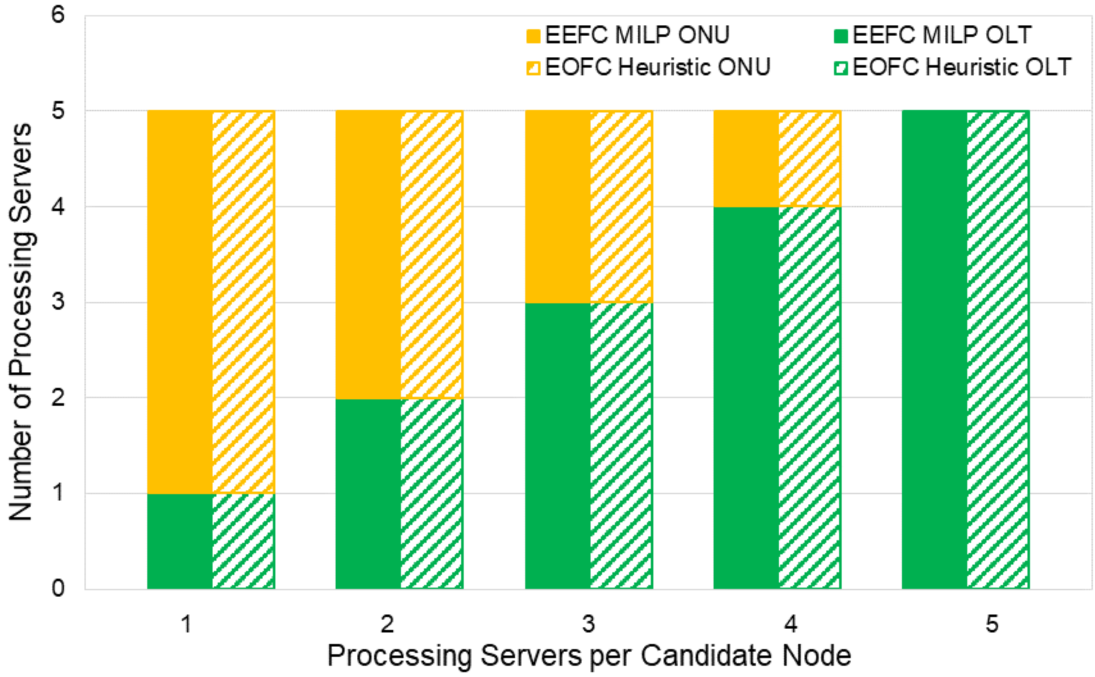


Figure 5.4: Optimal location of processing servers for EEFC model and EOFC heuristic for different numbers of processing servers per candidate node when 20% of patients can be served in a single processing server

Table 5.7: Number of candidate nodes utilised to place the processing servers

Type of Data	Number of Processing Servers per Candidate Node				
	1	2	3	4	5
Number of candidate nodes used to place the processing servers	5	4	3	2	1

Table 5.8: Optimisation gaps between the EEFC model and the EOFC heuristic for different numbers of processing servers per candidate node

	Gap (%)				
Number of processing servers per candidate node	1	2	3	4	5
Total energy	0.99%	0.97%	0.97%	0.97%	0.98%
Network energy	3.28%	3.09%	3.12%	3.15%	3.19%
Processing energy	0%	0%	0%	0%	0%

Figures 5.3 and 5.4 show the total energy consumption of the networking equipment and the processing and the optimal location to place the processing servers, respectively, for the EEFC model and the EOFC heuristic. Meanwhile, Tables 5.7 and 5.8 show the total number of candidate nodes utilised to place the processing servers and the optimisation gaps between the EEFC model and the EOFC heuristic, respectively. The results are shown for an increasing number of processing servers per candidate node. The results presented in Figure 5.3 revealed that the total energy consumption of both the networking equipment and the processing increased when the number of processing servers per candidate node increased from 1 to 2. The increasing energy consumption was attributed to the utilisation of the Ethernet switches dedicated to the health applications to connect multiple processing servers to the ONU and the OLT.

Figure 5.3 also shows that the total energy consumption of both the networking equipment and the processing slightly decreased when the number of processing servers per candidate node increased from 2 to 5. This was because limiting the number of processing servers per candidate node required the placement of the servers in multiple locations (i.e. OLT and

ONUs) as opposed to the optimal location at the OLT when a node could accommodate numerous servers. This is shown in Figure 5.4 where more processing servers are served by the OLT while reducing the number of utilised ONUs to place the processing servers when the number of processing servers per candidate node increases. Note that the larger the number of candidate nodes used to place the processing server is, the higher is the energy consumption because of the increasing amount of networking equipment (i.e. Ethernet switches) utilised.

Further, note that the data rate available per patient to transmit the analysed health data feedback traffic and the analysed health data storage traffic to the clinics and cloud storage, respectively, increased as fewer patients would be served at a node. Hence, more time could be allocated to send the video signal to the processing server. However, as the provisioned data rate to send the video signal was based on the number of PRBs, the same data rate and transmission time were used to send the video signal to the processing server (i.e. irrespective of the number of processing servers that could be served at each candidate node), as shown in Table 5.3, which resulted in utilising the same number of base stations to serve all the patients. Note that the increasing energy due to the increasing amount of networking equipment used to place the processing servers (i.e. Ethernet switches) with a small number of processing servers that could be served at each candidate node dominated the reduction in energy because of the low utilisation time of the network devices and the processing servers used to transmit the analysed health data feedback traffic and the analysed health data storage traffic to the clinics and the cloud storage, respectively.

Meanwhile, the results shown in Figure 5.3 also revealed that the total energy consumption of both the networking equipment and the processing of the EOFC heuristic approached those of the EEFC model. The overall optimisation gaps between the EOFC heuristic and the EEFC model for all numbers of processing servers per candidate node was less than 1%, as shown in Table 5.8. This difference was only due to the large amount of networking equipment (i.e. base stations) utilised in the EOFC heuristic as compared to the EEFC model. However, the increasing energy consumption of the networking equipment in the EOFC heuristic as compared to the EEFC model for all numbers of processing servers per candidate node was less 3.3%, as shown in Table 5.8. Also, the running times of the EOFC heuristic on a normal computer with 3.2 GHz and 16 GB RAM took 52 minutes to finish which is lower than the running time of the MILP on a high-performance computing (HPC) cluster with a 12 core and 64 GB RAM that took 8 hours to finish, when only one processing server can be served at each candidate node.

5.5 Conclusions

In this chapter, the impact of integrating fog computing at the network edge to serve a fall monitoring application that required a high data rate on the energy consumption of the networking equipment and the processing were investigated. The GPON architectures with fog computing were considered to monitor elderly patients who experienced falls and suffered from heart disease. Two scenarios related to the number of patients that could be served at each processing server and the number of processing servers allowed at each candidate node were considered to evaluate the performance of the fog computing in terms of the energy consumption of both the networking equipment and the processing. The EEFC model and the EOFC heuristic with additional parameters and constraints to meet the scenarios under consideration were utilised to optimise the number and the location of the processing servers at the network edge. Meanwhile, the EECC model was used as a benchmark to evaluate the performance of the EEFC model for the first considered scenario in terms of the total energy consumption of the networking equipment and the processing. The results revealed that when the number of processing servers allowed at each candidate node was not limited, there was 37.7% energy saving in the EEFC model as compared to the EECC model when 20% of the patients were served by a single processing server. The results also revealed that increasing the number of patients to be served by a single processing server reduced the total energy consumption of both the EEFC model and the EECC model because of the reduction of the number of utilised processing servers. The total energy saving in the EEFC model as compared to that in the EECC model increased to 52.2% when all the patients could be served by a single processing server. Meanwhile, the results also revealed that increasing the number of processing servers at each candidate node reduced the total energy consumption of the networking equipment and the processing. This reduction in energy was attributed to that fact that allowing more processing servers at each candidate node reduced the total amount of networking equipment (i.e. Ethernet switches) utilised where more processing servers were served by the OLT while reducing the number of the utilised ONUs to place the remaining processing servers that could not be served by the OLT. Furthermore, the performance of the EOFC heuristic for fall monitoring applications was evaluated. The results showed that the performance of the EOFC heuristic approached that of the EEFC model for the considered scenario 1 and scenario 2 with less than 2% and 1% total optimisation gap, respectively.

Chapter 6

Resilient infrastructure for health monitoring applications

6.1 Introduction

There are many approaches that have been introduced to improve service resilience at the cloud networking infrastructure as surveyed in [146] which range from designing and operating the facilities, servers, networks, to their integration and virtualisation. In [147], the utilisation of backup servers has been improved by 40% by using virtualisation which allows the sharing of backup servers in the geo-distributed data centres. However, the shared protection scheme proposed in this work requires high reserved bandwidth and can increase the latency of the secondary path between the primary and backup servers. Meanwhile, the work in [148] studies the benefits of relocation the primary and backup servers in terms of the total cost of both servers and network capacity. The study reveals that with the relocation, the cost of both servers and links capacity is reduced when considering protection against single link failures. Furthermore, the benefits of relocation are more pronounced for sparser topologies. The consideration of fog computing to perform local processing at the edge network, also improves services resilience. This has been studied in [46], whereby based on their simulation, fog computing can improve network resilience by offering local processing at the network edge and provides better response time compared to a cloud-only architecture, especially for an interactive request. However, to the best of our knowledge, no work has focused on reducing the energy consumption of both networking equipment and processing while improving the service resilience considering the server and network protection at the fog networking infrastructure level. Therefore, this chapter proposed a resilient energy efficient fog computing infrastructure for health monitoring applications. To improve the resilience of health monitoring applications, a 1+1 server protection scheme is considered. In this scheme, two servers, a primary and a secondary processing server are used to serve the health monitoring applications concurrently. The infrastructure is designed to be resilient against server failure under two scenarios related to the geographic location of primary and secondary servers and resilient against both server and network failures. Hence offering three levels of resilience. The first scenario considers the protection of servers without geographical constraint. In this scenario, the primary and secondary processing servers can be placed at the same node. The second scenario considers the protection of servers with geographical constraints, where the primary and secondary processing servers are not allowed to be placed at the same node. The latter offers higher levels of resilience compared to the former. This is because node failures at a given location are not improbable. The third scenario considers the protection of servers with geographical constraints and the protection of the network with disjoint links and nodes, offering a scenario with higher levels of resilience. In this scenario, the primary and secondary processing servers are not allowed to be placed at the same node, and the links and nodes used to transmit the data to and from primary and secondary processing servers are disjoint, as node and link failures in the network are not improbable. We consider the disjoint links and nodes only at the access layer, as the processing servers can only be placed at the access layer. A Mixed Integer Linear Programming (MILP) model, is used to optimise the number and locations of both primary and secondary processing servers so that the energy consumption of the networking equipment and processing are minimised. Also, a heuristic is developed for each scenario considered, for real-time implementation. The performance of the proposed resilient architecture is investigated, in terms of energy consumption of both networking equipment and processing under Electrocardiogram (ECG) and fall monitoring applications separately. For each application, the patients will send the necessary data to the primary and secondary processing servers for processing, analysis and decision making. Also, the energy penalty of increasing the level of resilience in the network is analysed.

6.2 The proposed resilient fog computing architecture for health monitoring applications

The architecture for the resilient fog computing infrastructure for healthcare applications over the Gigabit Passive Optical Networks (GPON) network consists of four layers, as shown in Figure 6.1. The first layer is the Internet of Things (IoT) layer that comprises of IoT devices to monitor the health of the patients and to send data to the network. The second layer is the access layer, which aggregates the health data from layer 1 using Long Term Evolution for Machines (LTE-M) base stations. Each base station is connected to an Optical Network Unit (ONU), and all the ONUs are connected to a single Optical Line Terminal (OLT). This layer is divided into several clusters, to improve the scalability and availability of the IoT system in vast networks [149]. Each cluster has one OLT connected to the ONUs of that cluster, and fog computing processing resources that can process and analyse health data, are available only at the ONUs and OLTs. Although the fog computing processing servers belong to a specific cluster, they can process the health data from any other clusters.

Placing the processing servers (PSs) at the ONU can reduce the energy consumption of networking equipment as such processing units are closer to the patients than cloud processing, but this can increase the required number of processing servers. Meanwhile, utilising the processing servers at the OLTs, reduces the required amount of processing servers as the OLT is a shared point between the base stations in the cluster, but this choice will increase the energy consumption of networking equipment [150].

The third layer is the metro layer, which aggregates and forwards data between the processing servers in the access network, while the fourth layer is the core layer, that is integrated with the central cloud and is used to store the analysed health data permanently.

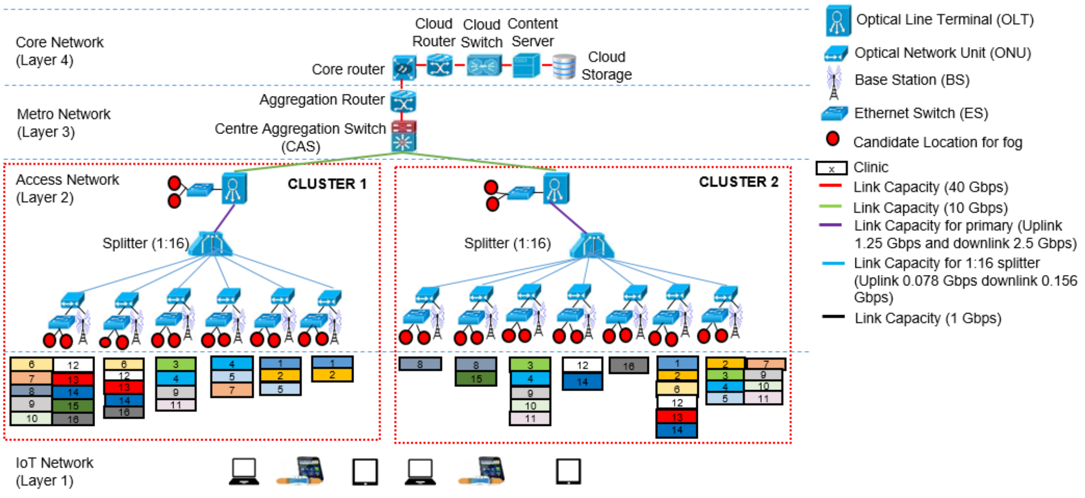


Figure 6.1: The resilient fog computing infrastructure for health monitoring applications

6.3 Mathematical model for energy efficiency and resilient infrastructure for fog computing considering server protection

In this section, the mathematical model for the two resilience scenarios related to the geographic location is provided. The MILP model with the objective function, to minimise the total energy consumption of both networking equipment and processing of the two resilience scenarios is provided.

6.3.1 Protection for servers without geographical constraints

To model the energy consumption minimised approach considering server protection without geographical constraints, the same sets, parameters, variables and objective function are utilised as in Section 4.6 and Section 5.3. We furthermore introduce additional variables as in Table 6.1 (also can be found in Appendix 1). Note that, as each candidate node can serve more than one processing server, hence ϕ_d is a variable.

Table 6.1: Additional variables used in MILP

Variables	
ωa_{sd}	Number of patients from clinic s served by primary processing servers located at candidate node d, $s \in CL, d \in FN$
ωb_{sd}	Number of patients from clinic s served by secondary processing servers located at candidate node d, $s \in CL, d \in FN$
Ya_d	$Ya_d = 1,$ if one or more primary processing servers are located at candidate node $d,$ otherwise $Ya_d = 0$, $d \in FN$

Yb_d	$Yb_d = 1$, if one or more secondary processing servers are placed at candidate node d, otherwise $Yb_d = 0$, $d \in FN$
z_d	z_d is a dummy variable that takes a value of $Ya_d \oplus Yb_d$, where $\oplus$ is an XOR operation, $d \in FN$
ϕa_d	Number of primary processing servers placed at candidate node d, $d \in FN$
ϕb_d	Number of secondary processing servers placed at candidate node d, $d \in FN$
τpa_d	Processing and analysis time of primary processing server (seconds) at candidate node d, $d \in FN$
τpb_d	Processing and analysis time of secondary processing server (seconds) at candidate node d, $d \in FN$

The model starts by defining the energy consumption of network (i.e. access, metro and core) and processing servers:

a) Energy consumption of access network, EAN:

The energy consumption of the access network, EAN is composed of the Long Term Evaluation (LTE) base stations', ONUs', and OLTs' energy consumption. The energy consumption of the LTE base stations, ONUs and OLTs are as given in Equation (3-4) - (3-14).

b) Energy consumption of metro network, EMN

The energy consumption of the metro network, EMN is composed of energy consumption of centre aggregation switches and aggregation routers. Note that the aggregation routers are only used to relay the analysed health data storage traffic as the candidate locations of the processing server are at the access layer as explained in Section 3.5. Hence, the raw health data traffic and analysed health data feedback traffic does not traverse the aggregation routers. Meanwhile, the centre aggregation switches are used to relay the raw health data traffic and analysed health data feedback traffic, between different clusters besides relaying the analysed health data storage traffic. Therefore, the energy consumption of metro network is given as:

$$EMN = (ECASP + ECASF + ECASS + EARS)\,\eta \qquad (6\text{-}1)$$

where the energy consumption of centre aggregation switches are needed to relay the raw health data, $ECASP$, analysed health data for feedback traffic, $ECASF$ and analysed health data for storage traffic, $ECASS$ are as given in Equation (3-63), Equation (3-64) and Equation (3-16), respectively. Meanwhile, the calculation of the energy consumption of aggregation routers needed to relay the analysed health data storage traffic, $EARS$ are as given in Equation (3-17).

c) Energy consumption of core network, ECN

The energy consumption of core network, ECN is composed of energy of the core routers. The energy consumption of core network and energy consumption of core routers are as given in Equation (3-18) and Equation (3-19), respectively.

d) Energy consumption of cloud, ECL

The energy consumption of cloud, ECL is composed of energy consumption of cloud routers, cloud switches, content servers and cloud storage. As explained in Section 3.5, the cloud storage is used to perform the storage task while other devices are only used to relay the analysed health data storage traffic. The energy consumption of the cloud is the same as given in Equation (3-20) while the energy of cloud routers, cloud switches and content servers are as given in Equation (3-21) - (3.23), respectively. Note that, for cloud storage, only one analysed health data storage traffic from both primary and secondary processing servers are stored. Therefore, the total analysed health data storage traffic S_i is divided by 2. The energy consumption of cloud storages is given in Equation (6-2):

$$ECSTS = 2 \sum_{i \in CST} \left(ICST \ x \ \zeta c_i + \frac{S_i}{2} \ \tau c \ \frac{PCST - ICST}{CCST} \right) \tau c \qquad (6\text{-}2)$$

The energy consumption of cloud storage is calculated, based on the size of analysed health data stored in the cloud storage and the time that the device is utilised.

e) Energy consumption of fog node, EFN:

The energy consumed by the fog, EFN reflects the energy consumed by primary and secondary processing servers and the energy consumed by the Ethernet switches. The energy consumption of the fog node is given as in Equation (4-2) where:

$$EPS = \sum_{d \in FN} (IPS\ (\phi a_d + \phi b_d)\ (\tau a + \tau b + \tau c) + PPS\ (\tau pa_d + \tau pb_d\,) \tag{6-3}$$

while the energy consumption of the Ethernet switches, $ETES$, required to relay raw health data traffic, analysed health data feedback traffic and analysed health data storage traffic are as given in Equations (4-4) – (4-7). The energy consumption of processing servers is determined by considering the idle energy consumption of the processing servers (i.e. primary and secondary processing servers) and the energy consumed to perform the processing as shown in Equation (6-3). The idle energy consumption of the primary and secondary processing servers is calculated by considering the time to receive raw health data traffic from clinic, $\tau a,$ the time to transmit the analysed health data feedback traffic to clinics, τb, as well as the time to transmit the analysed health data storage traffic to cloud storage, τc. Also, as explained in Section 3.5, the processing servers always work at its full utilisation, hence maximum power is consumed to perform the processing and analysis of the health data. Therefore, the energy consumption due to processing and analysis of the primary and secondary processing servers, is determined by considering the time to perform the processing and analysis, τpa_d and τpb_d, respectively.

The following are the modified and additional constraints used in addition to the constraints in Chapter 3 and Chapter 4, to model the energy consumption minimised approach, considering server protection without geographical constraints:

1) Association of patients from clinics to the processing server.

$$\omega a_{sd} \leq Pt_s\, Ya_d \quad ; \quad \forall s \in CL, \forall d \in FN \tag{6-4}$$

$$\omega b_{sd} \leq Pt_s\, Yb_d \quad ; \quad \forall s \in CL, \forall d \in FN \tag{6-5}$$

Constraint (6-4) and constraint (6-5) replaced constraint (3-28) to allocated patients from clinic $s,$ to be served by the primary and secondary processing servers at fog located at node d, respectively. Note that, if a patient is allocated to a candidate location, this location should have fog.

$$\sum_{d \in FN} \omega a_{sd} = Pt_s \quad ; \quad \forall s \in CL \tag{6-6}$$

$$\sum_{d \in FN} \omega b_{sd} = Pt_s \quad ; \quad \forall s \in CL \tag{6-7}$$

Constraint (6-6) and constraint (6-7) replaced constraint (3-29), to ensure that all patients at clinic $s,$ are assigned to the primary and secondary processing servers located at any node d, respectively.

2) Traffic from clinics to processing servers.

$$P_{sd} = (\omega a_{sd} + \omega b_{sd})\, \delta a \quad ; \ s \in CL, d \in FN \tag{6-8}$$

Constraint (6-8) replaced constraint (3-30), to calculate the raw health data traffic from clinic s, to both primary and secondary processing servers located at node d. This is based on the association of patients from the clinic to the primary processing server, ωa_{sd}, the association of patients from clinic to secondary processing server, ωb_{sd} as well as the data rate provisioned for each patient, δa, to perform the transmission.

3) Traffic from processing server to clinics.

$$F_{sd} = (\omega a_{ds} + \omega b_{ds})\, \delta b \quad ; \ \forall s \in FN, d \in CL \tag{6-9}$$

Constraint (6-9) replaced Constraint (3-31), to calculate the analysed health data feedback traffic from primary and secondary processing servers, located at node s, to clinic d. This is based on the total number of patients in the clinic, served by the primary processing servers, ωa_{ds}, the total number of patients in the clinic served by the secondary processing servers, ωb_{ds}, and the data rate provisioned for each patient, δb, to perform the transmission.

4) Traffic from processing server to cloud storage.

$$S_{sd} = \sum_{i \in CL} (\omega a_{is} + \omega b_{is})\, \delta c\ \ \delta_{sd} \quad ; \ \forall s \in FN, d \in CST \qquad \text{(6-10)}$$

Constraint (6-10) replaced constraint (3-32), to calculate the analysed health data storage traffic from primary and secondary processing servers located at node s, to cloud storage d. This is based on the total number of patients in the clinic served by primary processing servers, ωa_{is}, the total number of patients in the clinic served by secondary processing servers, ωb_{is}, and the data rate provisioned for each patient, δc, to perform the transmission.

5) Nodes used to connect the servers.

$$Ya_d + Yb_d = 2\, Y_d - z_d \quad ; \forall d \in FN \qquad \text{(6-11)}$$

$$\phi a_d + \ \phi b_d \ \leq N \quad ; \ \forall \mathrm{d} \in FN \qquad \text{(6-12)}$$

Constraint (6-11), is to determine the nodes that are used to place the processing servers where $Y_d = 1$ if at least any of Ya_d and Yb_d are equal to 1 $(Ya_d + Yb_d)$, otherwise zero. This is achieved by, introducing a binary variable

z_d which is only equal to 1 if Ya_d and Yb_d are exclusively equal to 1 ($Ya_d \oplus Yb_d$), otherwise it is zero. Constraint (6-12) replaced constraint (5-2), to ensure that the number of processing servers at node d does not exceed the maximum number of processing servers allowed at each candidate node N.

6) Maximum number of patients served at each processing server.

$$\sum_{s \in CL} \omega\mathrm{a}_{sd} \leq \Omega max\ \phi a_d \quad ; \ \forall d \in FN \tag{6-13}$$

$$\sum_{s \in CL} \omega\mathrm{b}_{sd} \leq \Omega max\ \phi b_d \quad ; \ \forall d \in FN \tag{6-14}$$

Constraint (6-13) and constraint (6-14) replaced constraint (3-59), to ensure that the number of patients served by each primary and secondary processing server at node d, respectively does not exceed its maximum number of users, Ωmax. Note that, the model also allows more than one primary processing server, ϕa_d and secondary processing servers, ϕb_d, to be deployed at the same fog located at node d if the number of users is higher than Ωmax.

7) Processing and analysis time at each processing server.

$$\tau pa_d = \sum_{s \in CL} m\ \omega\mathrm{a}_{sd} + \acute{c}\ \phi a_d \quad ; \forall d \in FN \tag{6-15}$$

$$\tau pb_d = \sum_{s \in CL} m\ \omega \mathrm{b}_{sd} + ć\ \phi b_d \quad ; \forall d \in FN \tag{6-16}$$

Constraint (6-15) and constraint (6-16) replaced constraint (3-60), to calculate the processing and analysis time of the primary processing server and secondary processing server at node d, respectively. This is based on the total number of patients served by the processing server (i.e. $\omega \mathrm{a}_{sd}$ for primary processing server and $\omega \mathrm{b}_{sd}$ for secondary processing server) and number of processing servers used (i.e. ϕa_d for primary processing server and ϕb_d for secondary processing server), where m and $ć$ are constant value.

8) Storage capacity constraint at each processing server.

$$\sum_{s \in CL} \omega \mathrm{a}_{sd}\ \alpha \leq \Lambda max\ \phi a_d \quad ; \ \forall d \in FN \tag{6-17}$$

$$\sum_{s \in CL} \omega \mathrm{b}_{sd}\ \alpha \leq \Lambda max\ \phi b_d \quad ; \ \forall d \in FN \tag{6-18}$$

Constraint (6-17) and constraint (6-18) replaced constraint (3-61), to ensure that the storage capacity of a primary processing server and secondary processing server at node d, do not exceed its maximum capacity, Λmax, respectively. Note, that the model also allows more than one primary processing servers, ϕa_d and secondary processing servers, ϕb_d to be deployed at the same fog processing unit located at node $d,$ if the size of the data is higher than Λmax.

6.3.2 Protection for servers with geographical constraints

This section considered server protection with geographical constraints, where the primary and secondary processing servers are not allowed to be placed at the same node. Typically, most service providers place their primary and secondary services in distant locations, to increase resilience. For example, BackupVault, which is a leading provider of online cloud backup for businesses in United Kingdom, locate their primary data centre in Slough, UK; while the second data centre for redundancy is located in Reading, UK [151]. Therefore, this work considered that the nodes serving the primary processing servers are not allowed to serve any of the secondary processing servers. The same parameters, variables, constraints and objective functions in Section 6.3.1 are utilised. However, to ensure that the locations of both primary and secondary processing servers are different, constraint (6-11) is replaced with Equation (6-19), as shown below:

$$Ya_d + Yb_d = Y_d \; ; \; \forall d \in FN \tag{6-19}$$

where constraint (6-19), ensures that either primary or secondary processing servers can be placed at one location d as the maximum value for Y_d is 1.

6.4 Realistic parameter consideration for ECG monitoring applications

This section elaborates in detail on the methodologies for determining the considered model input parameters, such as network layout under GPON network, the number of monitored patients in West Leeds, UK, and the calculation of data rate for traffic transmission for ECG monitoring applications.

6.4.1 Network layout under GPON network and total number of monitored heart patients in West Leeds, UK

In this section, the number and location of both primary and secondary processing servers are optimised in the fog architecture, as described in Section 6.2. Also, the same location of clinics and LTE base stations and the same number of patients in the clinics as in Chapter 4, are considered. However, the complexity of the MILP model grows exponentially with the number of nodes in the network. Therefore, a scenario with 16 clinics that have a total of 300 patients and 13 LTE base stations, is considered using the locations at West Leeds as a case study. The 13 LTE base stations are selected, based on the nearest distance between the available base stations (BSs) and the clinics. Two clusters are considered as a case study and the clinics are connected to up to two of the nearest BSs in each cluster, as shown in Figure 6.1. For example, clinic 13, shown in red, is connected to two base stations in cluster 1, and also a single base station in cluster 2. Figure 6.1

shows the resilient fog computing architecture under the GPON network, while Table 6.2 presents the considered total number of patients at each clinic.

Table 6.2: Number of monitored patients in clinics

Clinic	Total Number of Patients
Craven Road Medical Practice	20
Leeds Student Practice	68
Hyde Park Surgery	13
Burton Croft Surgery	15
Laurel Bank Surgery	16
Kirkstall Lane Medical Centre	11
Burley Park Medical Centre	23
Thornton Medical Centre	18
Beech Tree Medical Centre	16
Hawthorn Surgery	4
Priory View Medical Centre	20
Abbey Grange Medical Centre	9
Vesper Road Surgery	16
The Highfield Medical Centre	25
Dr G Lees & Partners	10
Whitehall Surgery	16

6.4.2 Data rate calculation for traffic transmission in the network

As in Chapter 4, in this section, the same 30 seconds ECG recording signal (Π) with a size of 252.8 kbits, is utilised. Patients send their ECG signals to the network to be processed and analysed at both primary and secondary processing servers of the fog layer. The relationship between the processing and analysis time of the signal and the number of patients to perform the processing at both processing servers utilising the Pan-Tompkins algorithm, are retrieved from the experiment conducted in Section 4.2.3. Based on the results, the duration of processing and analysis at a given number of patients (Pat) is: $\tau p = 0.002\ Pat + 4.6857$.

In this work, the number of patients that can be served in a single processing server Pat is limited, in order to investigate the distribution of primary and secondary processing servers in the network, with increasing demands. Therefore, the maximum Pat that can be served at a single server is considered to be 20% of the total number of patients from the 16 clinics, which is the lowest demand evaluated in the network. Based on our experimental results, the size of the processed and analysed data α, was found to be 256 bits. This result will be sent from the primary and secondary processing servers to the cloud for permanent storage, but only one copy will be stored. The same principle applies to the data that is sent to the clinic from both servers.

The energy consumption of networking equipment and processing is calculated, based on the timing constraints set by the American Heart Association (AHA) [25]. As in Chapter 4, 4 minutes (i.e. $\tau t = 4\ \text{minutes}$) is considered as the maximum duration to save heart patients. The $4\ \text{minutes}$ include the time to record the 30-second ECG signal, τm, the time to transmit the raw ECG signals to both servers for processing task, τmax, the time for processing and analysis, τp, and the time to transmit the analysed ECG data for feedback, τb. To determine the available time to transmit the raw ECG signal to the processing servers, τmax, the time of both processing and analysis, τp is calculated based on the maximum number of patients that can be served by a single processing server (Pat) and the time to send the analysed ECG data to the clinics for feedback, τb while considering the 30 seconds of ECG recording, τm from the patient for τt equal 4 minutes.

The feedback time is calculated as follows; First determine the maximum number of patients ($MaxP$) that can be served by the processing servers at each candidate node. Due to the limited spaces at fog nodes and the complexity of the model to have more base stations, the maximum number of processing servers, N that can be connected at each candidate node is limited to 3, 4, 5, 6, 7 and 8 while each processing server can serve a maximum of Pat. Therefore, $MaxP = N\ Pat$. Then, the minimum shared capacity between the candidate locations of processing servers at the access layer to the Long Term Evaluation (LTE) base station (i.e. uplink between ONU and OLT) is determined. As the link capacity will be shared by the maximum number of patients, the processing servers can serve at a node, the link capacity is divided by $MaxP,$ to obtain the data rate for each patient to transmit the analysed data to the clinics (δf). The reason for limiting the feedback data rate by the data rate available for healthcare applications in GPON links, is as explained in Section 4.2.4. However, in LTE BS, each user is given a minimum of one physical resource block (PRB). Therefore, the minimum data rate that can be given to each patient is 336 bps per single PRB (r) when using Quadrature Phase Shift Keying (QPSK) modulation format as calculated in Section 4.2.4. Due to this, a maximum number of PRBs (REF) are allocated for each patient while ensuring that the given data rate does not exceed δf. Hence, the data rate for feedback is $\delta b = r\ REF$, while the feedback time is

$\tau b = \alpha / \delta b$. Also, a maximum number of PRBs are allocated to each patient to send their ECG signal to the processing servers, while ensuring that the given data rate, δa is higher than $\Pi / \tau max$ to ensure the system meets the 4 minutes deadline. Due to this, the transmission time to send the ECG signal to the processing servers for processing and analysis is $\tau a = \Pi / \delta a$. Note that, the link capacities are shared by multiple applications. Therefore, as explained in Section 3.2.1, 0.3% of the maximum available link capacity is considered for healthcare applications. The data rate for storage task (δc), is calculated by dividing the minimum shared uplink capacities between the candidate location of processing servers and the cloud storage (i.e. the link between the ONU and OLT) by $MaxP$. Hence, the transmission time to send the analysed ECG data to the cloud for permanent storage is $\tau c = \alpha / \delta c$.

There are five approaches considered with 20%, 40%, 60%, 80% and 100% of the total number of patients in the 16 clinics, to investigate the impact of increasing the number of patients on the energy consumption of networking equipment and processing. This was done while considering the two server protection scenarios and the different number of allowed processing servers at each candidate node. Table 6.3 shows the data rate and transmission time to transmit the raw ECG data to the processing server, to transmit the analysed ECG data to the clinics and cloud for feedback and permanent storage, respectively. This is shown for the different number of processing servers per candidate node, N. Note that, the data rate and the time to transmit the raw ECG data to the processing servers for each number of processing servers per candidate node are the same for all approaches. This is because the data rate given to each patient is based on the number of allocated PRBs while ensuring the total data rate provided by the total number of PRBs per patient, is equal to or higher than the minimum data rate required so that the system can work within 4-minutes. Therefore, the same amount of PRBs are given to each patient under the different number of processing servers per candidate node, although their required minimum data rate is different.

Table 6.3: Data rate and related time for a different number of processing servers per candidate node N, for ECG monitoring applications

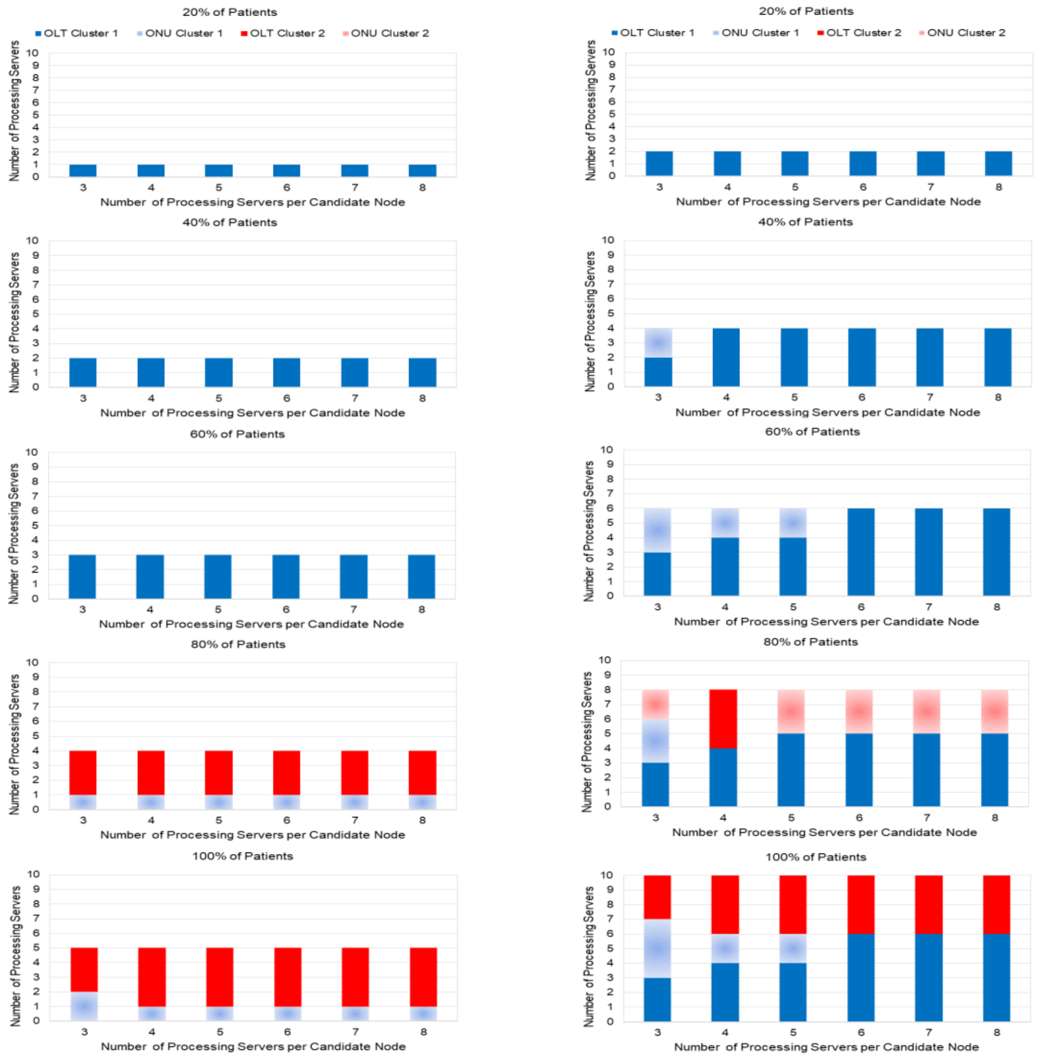
Type of Data	3 PSs	4 PSs	5 PSs	6 PSs	7 PSs	8 PSs
Data rate to transmit ECG signal to processing server, δa (kbps)	1.344	1.344	1.344	1.344	1.344	1.344
Transmission time to transmit ECG data to processing server, τa (s)	188.1	188.1	188.1	188.1	188.1	188.1
Data rate to transmit analysed ECG data to clinics, δb (kbps)	1.008	0.672	0.672	0.336	0.336	0.336
Transmission time to transmit analysed ECG data to clinics, τb (s)	0.254	0.381	0.381	0.762	0.762	0.762
Data rate to transmit analysed ECG data to cloud storage, δc (kbps)	1.28	0.96	0.768	0.64	0.548	0.48
Transmission time to transmit analysed ECG data to cloud storage, τc (s)	0.2	0.267	0.333	0.4	0.467	0.533

6.5 Results and analysis of the MILP model for ECG monitoring applications considering server protection

In this section, the impact of increasing the level of resilience for server protection to the energy consumption of networking equipment and processing, is investigated. The evaluation is divided into two steps. For each step, an analysis is carried out to determine (i) the locations of the processing servers in each scenario, (ii) the energy consumption of the networking equipment, (iii) the energy penalty of networking equipment due to the increasing level of resilience. The first step is comparing the non-resilient scenario with a resilient scenario, without geographical constraint. Secondly, it is comparing the resilient scenario without geographical constraint with the more resilient scenario considering geographical constraints. The same power profile of equipment and the same networking devices used in Chapter 4 in Table 4.3 and Table 4.8, are considered to evaluate the energy consumption of the networking equipment and processing. Also, as in the previous chapters, the networking devices are shared by multiple applications. Thus, only 0.3% of the idle power is considered, while the processing server and Ethernet switch are dedicated to the healthcare applications. Note that, as in the previous chapters, the AMPL software with CPLEX 12.8 solver running on high-performance computing (HPC) cluster with a 12 core CPU and 64 GB RAM was used as the platform for solving the MILP models.

6.5.1 Performance analysis of server protection resilient scenario without geographical constraints

In this section, the performance of the non-resilient scenario is used as a benchmark to evaluate the resilient scenario without geographical constraints in terms of the energy consumption of networking equipment and processing for ECG monitoring applications.



(a) (b)

Figure 6.2: Optimal location of processing servers for (a) non-resilient scenario and (b) resilient scenario without geographical constraints for ECG monitoring applications

The results in Figure 6.2, show that the number of processing servers for the resilient scenario are double that of the non-resilient scenario. This is because, the non-resilient scenario only has primary processing servers while the resilient scenario consists of a secondary processing server for each primary processing server, for server protection purposes. The results show that increasing the percentage of patients, has resulted in increasing the number of processing servers. For the non-resilient scenario, at demand level of 20%, 40%, 60%, 80% and 100%, the number of processing servers required to serve all patients are one, two, three, four and five, respectively. Meanwhile, for the resilient scenario, at demand levels of 20%, 40%, 60%, 80% and 100%, the total number of required processing servers (i.e. primary and secondary processing servers) are two, four, six, eight and ten, respectively.

The results, also show that the OLT is always chosen to place the processing servers as it is the nearest shared point to the patients (i.e. the OLT is connected to all base stations of the same cluster) which reduces the number of required processing servers and the number of hops to transmits the ECG signal to the processing servers.

The processing servers are placed at only one cluster when the percentage of patients considered in the network is equal to or less than 60% as shown in Figure 6.2-(a) and Figure 6.2-(b). This is because, all patients can be served by the base stations in one cluster only. Therefore, for the resilient scenario without geographical constraints, to reduce the number of utilised networking equipment in the network, the ONU is selected to place the remaining processing servers, which cannot be allocated at the OLT at the same cluster, while for the non-resilient scenario the processing servers are only placed at the OLT.

However, increasing the percentage of patients to 80% and 100% has resulted in utilising the BSs, ONUs and OLTs in both clusters. For the non-resilient scenario, the primary processing servers are placed at the OLT and ONU of different cluster when the demand increases to 80% and 100%. This is because, the OLT does not have enough capacity to support all of the traffic. The OLT of cluster 2 is occupied first, and the remaining demands are sent to the ONU of the cluster 1, to reduce the total amount of data traversing the network as ONUs are directly connected to the patients. For the resilient scenario without geographical constraints, at a demand level of 80%, and three PSs available at each candidate node, the OLT and ONUs of cluster 1 are occupied first, and the remaining demands are sent to the ONU of cluster 2. This is due to the same reason, as explained for the non-resilient scenario. However, when the demand level increases to 100%, the OLTs of both clusters and only the ONU of one cluster are used. The model did not use multiple ONUs to place the processing server to reduce the number of utilised Ethernet switches. Also, when five processing servers are allowed at each candidate node, the processing servers are placing at both OLTs and one ONU. This is mainly to reduce the number of utilised base stations, as the base station consumed more energy than the Ethernet switch.

Figure 6.2 also shows that increasing the number of processing servers that are available at each candidate node can also reduce the number of candidate nodes to place the processing servers for the resilient scenario without geographical constraint. For instance, the number of candidate nodes to place the processing servers reduces when the number of processing servers per node increases to four, at demand levels of 40%, 80% and 100% and six for the considered demands of 60% and 100%.

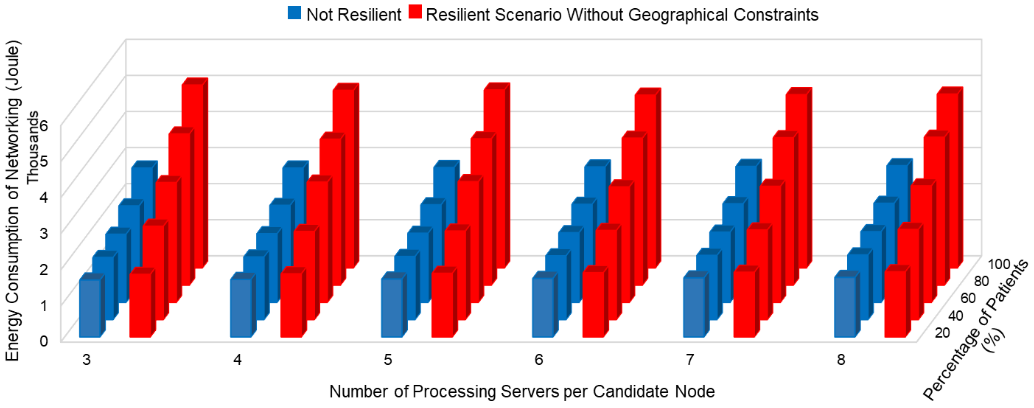


Figure 6.3: Energy consumption of networking equipment for non-resilient scenario and resilient scenario, without geographical constraints for ECG monitoring applications

The results in Figure 6.3, show that energy consumption of networking equipment increases as the demand increases for both scenarios. However, the increasing rate of energy consumption of networking equipment due to the increasing demand for the resilient scenario, is higher than the non-resilient scenario. The results also show that the energy consumption of networking equipment of the resilient scenario without geographical constraints, is always higher than the non-resilient scenario for all levels of demand and number of processing servers per node. This is because, the total traffic traversing the networking equipment for the resilient scenario is double compared to the non-resilient scenario, hence increasing the total number of utilised networking equipment. This increase in energy consumption is one of the key penalties for having resilience.

Figure 6.3 also shows that at a demand level equal to or more than 40%, the energy consumption of networking equipment of the resilient scenario reduced significantly when the number of processing servers increased from three to eight. This is because, increasing the number of processing servers per candidate node has resulted in placing the processing servers at their optimal locations besides reducing the number of utilised nodes.

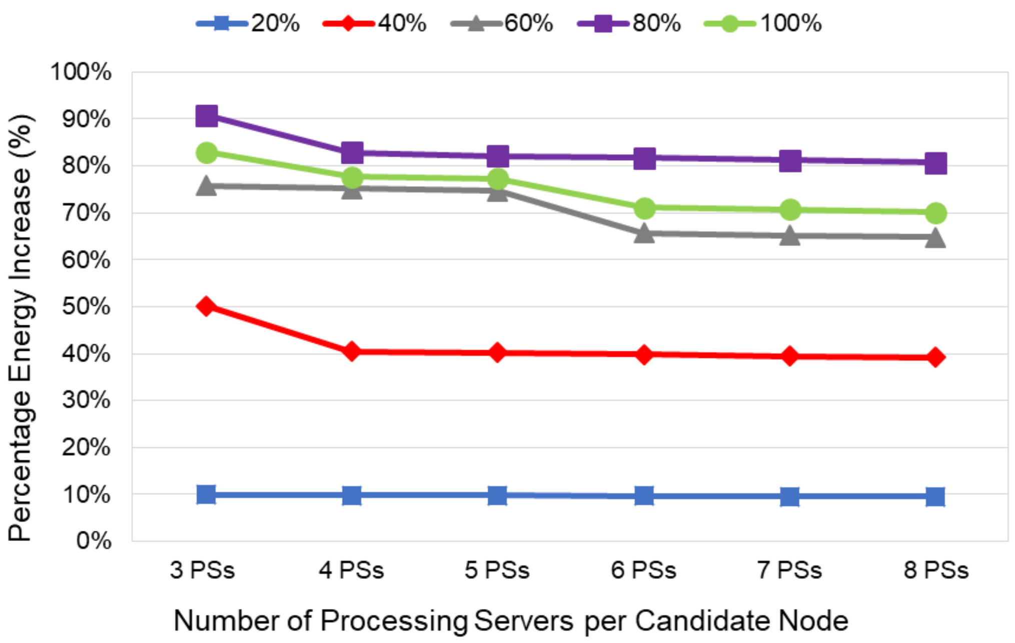


Figure 6.4: Percentage of energy penalty of networking equipment for resilient scenario, without geographical constraints compared to non-resilient scenario for ECG monitoring applications.

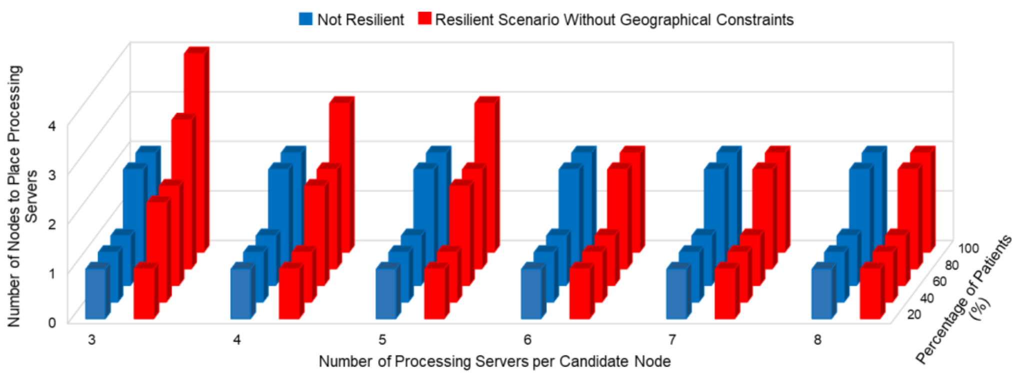
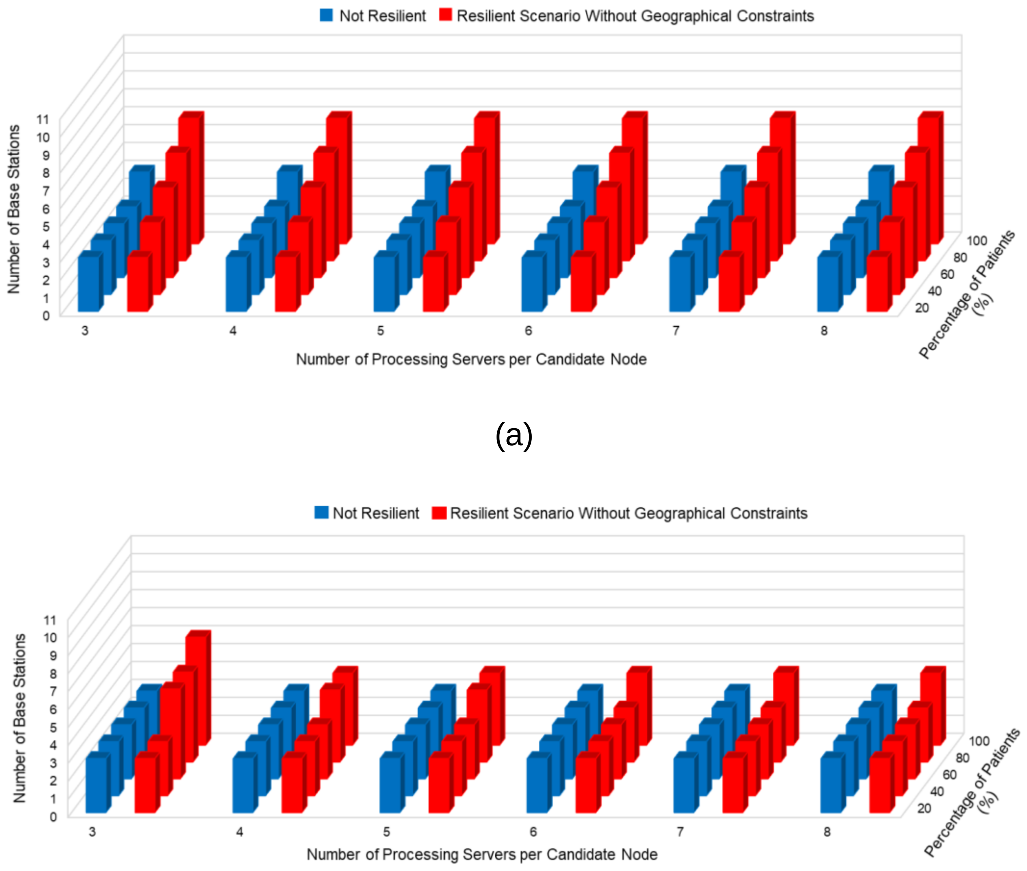


Figure 6.5: Number of candidate nodes used to place processing servers for non-resilient scenario and the resilient scenario, without geographical constraints for ECG monitoring applications.



(a)

(b)

Figure 6.6: Number of base stations used to send (a) the raw ECG signal for processing and (b) the analysed ECG signal for feedback, for non-resilient scenario and resilient scenario without geographical constraints under different percentages of patients and number of processing servers per candidate node, for ECG monitoring applications

The results in Figure 6.4, show that the energy penalty (defined as the difference in energy consumption between the resilient and the non-resilient cases) increases when the level of demand increases from 20% to 80%. This is because, at demand levels of 20% to 80%, the number of utilised base stations to serve all patients to send their ECG signal to the processing servers for the non-resilient scenario are the same while for the resilient scenario, the number of base stations increases with the increasing demand, as shown in Figure 6.6-(a). The increasing number of base stations under the resilient scenario is because, each patient will send two ECG signals to both primary and secondary processing servers, hence requiring a high number of base stations to serve all patients and this number increases as the demand increases.

For the non-resilient scenario, each patient only sends one ECG signal to the primary processing servers, and the same number of base stations are used, as they can accommodate the increasing demand by up to 80%. However, at a demand level of 100%, the energy penalty is lower than 80%. This is because, at a demand level of 100%, the number of base stations used for the non-resilient scenario increases, hence increasing the energy consumption of networking equipment of the non-resilient scenario. Figure 6.4, also shows that increasing the number of processing servers at each candidate node can significantly reduce the energy penalty when the demand is equal to or is higher than 40%. This is due to the reduction in the number of candidate nodes used to place the processing servers for the resilient scenario as shown in Figure 6.5, where more processing servers can be placed at the same node when the number of processing servers allowed at each candidate node increases.

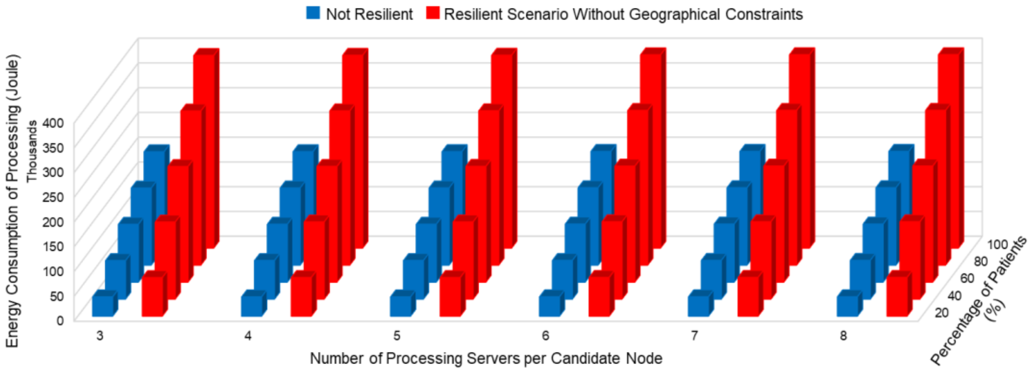


Figure 6.7: Energy consumption of processing for non-resilient scenario and resilient scenario, without geographical constraints for ECG monitoring applications

The results in Figure 6.7, show that the energy consumption of processing for the resilient scenario is higher than the non-resilient scenario. This is because the number of processing servers for the resilient scenario, is double that of the non-resilient scenario. The results also show that the energy consumption of processing increases, as the demands increase for both scenarios. This is because increasing the number of patients increases the number of processing servers proportionally.

However, the same total number of servers is used in both scenarios under constraints on the number of processing servers per candidate node, as the patients were optimally consolidated in the servers. Also, for both scenarios, there is a slight increase in energy consumption, when more processing servers are allowed per candidate node. The increase is due to the increasing utilisation time of the processing servers to send the feedback and storage traffic, with the increasing number of processing servers per candidate node, as shown in Table 6.3.

6.5.2 Performance analysis of resilient scenario considering geographical constraints

In this section, the performance of the resilient scenario without geographical constraints is used as a benchmark to evaluate the increasing level of resilience gained by considering the geographical constraints, in terms of the energy consumption of networking equipment and processing.

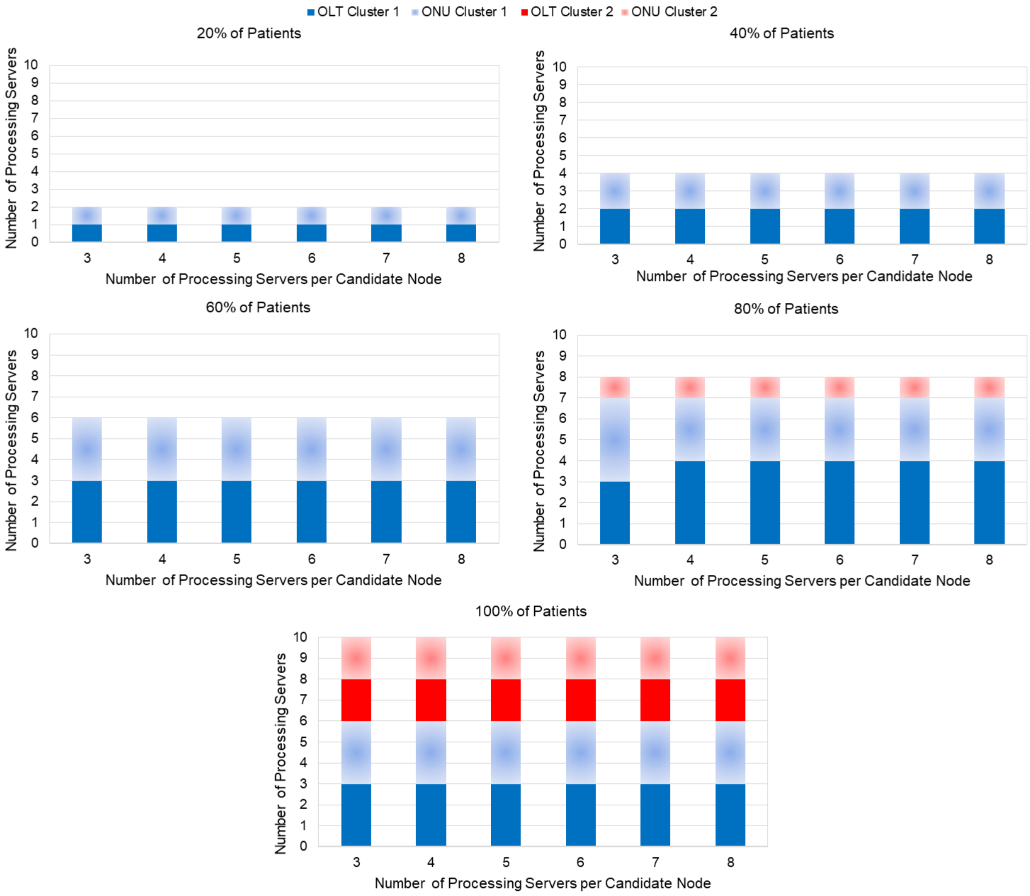


Figure 6.8: Optimal location of processing servers for resilient scenario, considering geographical constraints for ECG monitoring applications

The results in Figure 6.8, show that the OLT is always used to place the processing servers as in the previous scenarios. The results also show that the processing servers are placed at only one cluster, when the percentage of patients is equal to or less than 60%. This is to reduce the utilisation of the networking equipment. However, due to the geographical constraints, at least two locations are required to place the primary and secondary processing servers. Therefore, both OLT and ONU of the same clusters are selected to place the processing servers, separately.

Figure 6.8 also shows that at a high level of demand (i.e. 80% and 100%), the BSs, ONUs and OLTs from both clusters are utilised. The results show that at a demand level of 80% for all processing servers per candidate node, the OLT and ONUs of cluster 1 are occupied first, and due to the limited number of resources of the base stations in cluster 1 to serve the patients, the remaining demand is sent to the ONU of cluster 2. This is to reduce the total amount of data traversing the network as ONUs are directly connected to the patients. The results also show that, at the demand level of 80%, when the number of processing servers allowed at each node increases to four, the number of utilised nodes to place the processing servers are reduced as more processing servers are placed at the OLT.

However, when the demand level increases to 100%, the OLT and the ONU of both clusters are used to accommodate the increasing number of processing servers in the network for all processing servers per candidate node. The results show that increasing the number of processing servers per candidate node does not affect the location of placing the processing servers, as optimal locations are selected.

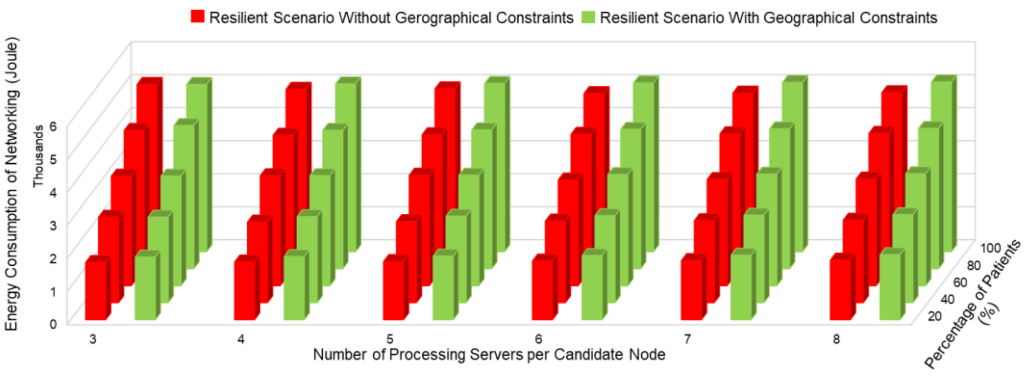


Figure 6.9: Energy consumption of networking equipment for resilient scenario, without geographical constraints and resilient scenario considering geographical constraints for ECG monitoring applications

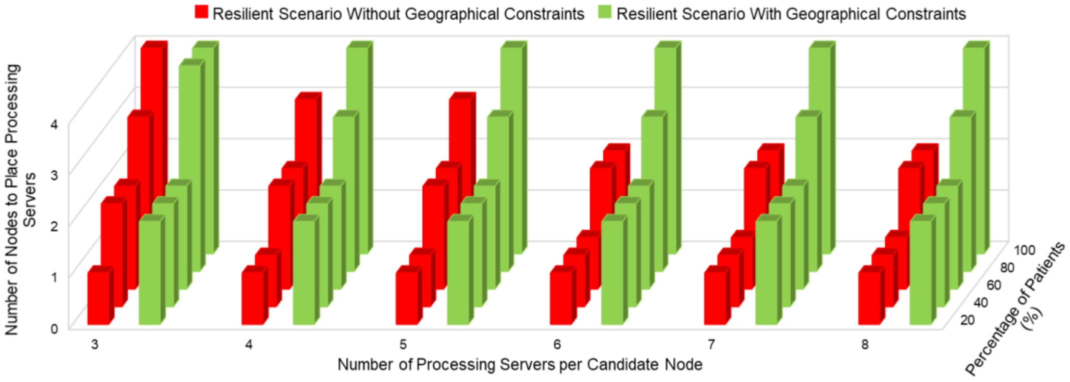
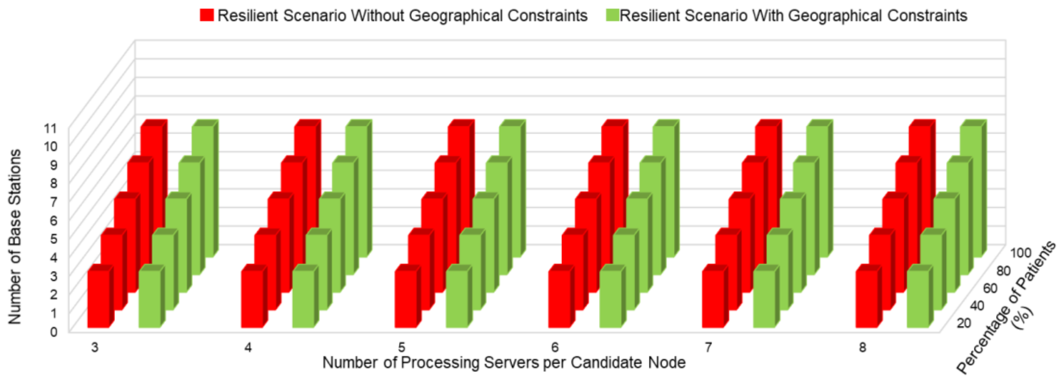
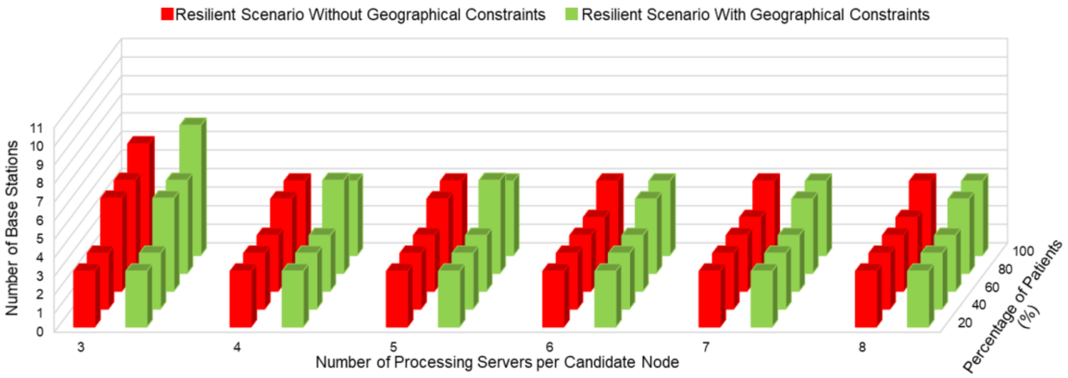


Figure 6.10: Number of candidate nodes used to place processing servers for resilient scenario, without geographical constraints and resilient scenario considering geographical constraints for ECG monitoring applications.



(a)



(b)

Figure 6.11: Number of base stations used to send (a) the raw ECG signal for processing and (b) the analysed ECG signal for feedback, for resilient scenario without geographical constraints and resilient scenario, with geographical constraints under different percentages of patients and number of processing servers per candidate node, for ECG monitoring applications

The results in Figure 6.9 show that, at demand levels of 40%, 60% and 100% and when the number of available processing servers is 3, the energy consumption of networking equipment for both scenarios is the same. This is due to the same number of utilised networking equipment, where the same number of base stations are used to serve the patients to send their ECG signal to the processing servers, and the same number and location of nodes are used to place the processing servers for both scenarios, as shown in Figure 6.11 and Figure 6.10, respectively.

However, at a demand level of 60% and when four and five processing servers are allowed at each candidate node, the energy consumption of networking equipment with the more resilient scenario is slightly higher than the resilient scenario without geographical constraints, although the same number of base stations and nodes are used to place the servers for both scenarios. This is due to the different locations of the processing servers in the network for both scenarios where for the more resilient scenario, the location of processing servers has led to more data traversing the networking equipment than the resilient scenario, without geographical constraints.

Meanwhile, for the other levels of demand and number of processing servers per candidate node, the energy consumption of the more resilient scenario is higher than the resilient scenario without geographical constraints, as shown in Figure 6.9. This is because, considering the geographical constraint increases the total number of utilised nodes to place the processing servers, as shown in Figure 6.10. Hence, the number of utilised networking equipment in the more resilient scenario increases. This increase in energy consumption, is the penalty for having a higher level of resilience.

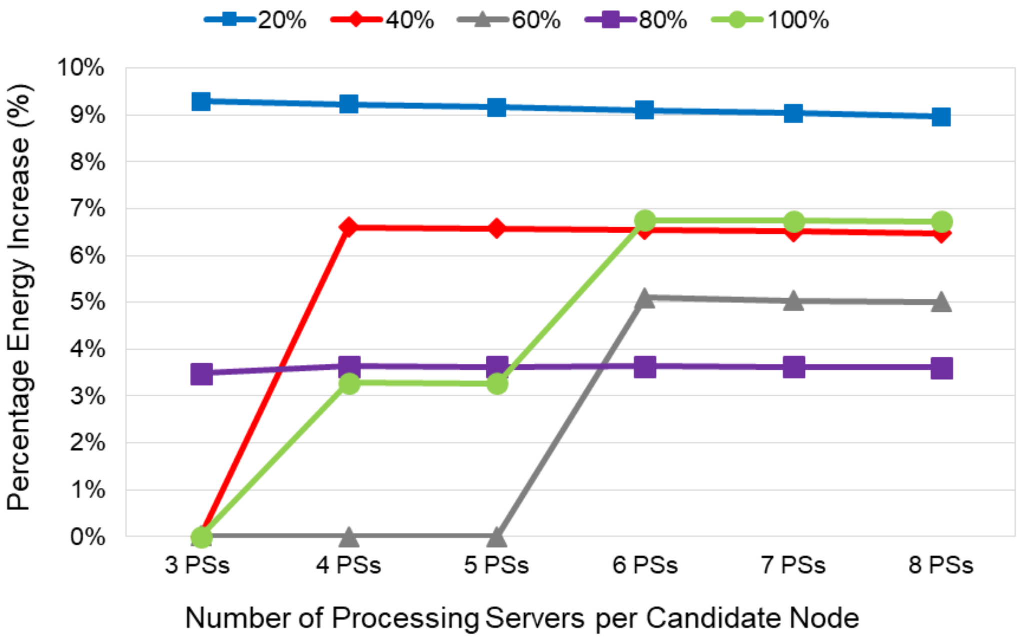


Figure 6.12: Percentage of energy penalty of networking equipment for the resilient scenario, considering geographical constraints, compared to the resilient scenario without geographical constraints for ECG monitoring applications

The results in Figure 6.12, show that increasing the level of resilience to consider geographical constraints, does not incur any energy penalty at demand levels of 40%, 60% and 100%; when three processing servers are available at each node. This is due to the same number of utilised networking equipment in both scenarios (i.e. nodes to place the processing servers and base stations to send the processing traffic). However, at demand levels of 20% and 80%, increasing the level of resilience to consider geographical constraints has resulted in an energy penalty. This is because, at these specific demands, a higher number of candidate nodes are used to place the processing servers for the more resilient scenario, compared to the resilient scenario without geographical constraints, as shown in Figure 6.10.

Figure 6.12 also shows that increasing the number of allowed processing servers at each candidate node can increase the energy penalty when the demand is 40%, 60%, 80% and 100%. The increase in energy penalty is due to the decreasing number of candidate nodes available to place the processing servers with a resilient scenario without geographical constraints, as shown in Figure 6.10. However, at demand levels of 20% and 80%, increasing the number of processing servers per candidate node, does not result in significant impact on the energy penalty. This is because, at this specific demand, the same number of candidate nodes are used to place the processing servers and the same number of base stations are used to send the ECG signal to the processing servers in both scenarios, as shown in Figure 6.10 and Figure 6.11, respectively.

Figure 6.12 also shows that, when the number of processing servers allowed at each node is equal to or higher than 6, the energy penalty decreases as the demand increases from 20% to 80%. This is because the same number of base stations are used in both scenarios to send the ECG signal to the processing servers, as shown in Figure 6.11-(a). However, the energy penalty at a demand level of 100% is higher than 40%, as the number of candidate nodes used to place the processing servers for the more resilient scenario is doubled, in comparison to the resilient scenario without geographical constraints, as shown in Figure 6.10.

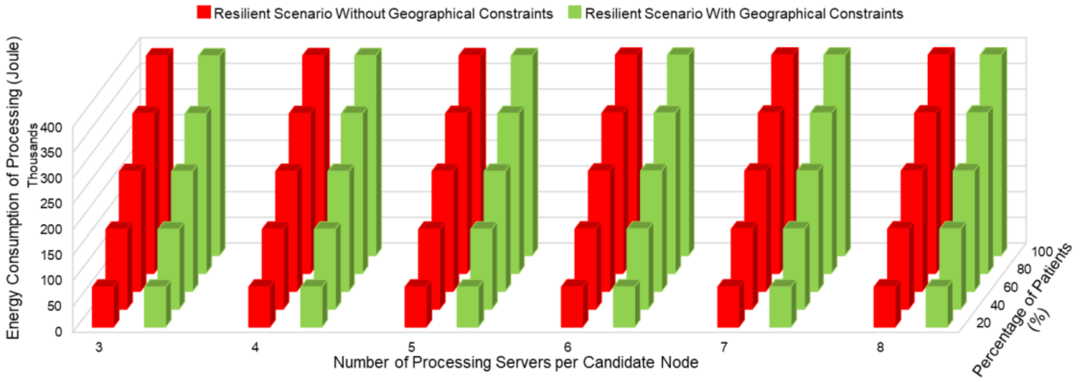


Figure 6.13: Energy consumption of processing for resilient scenario without geographical constraints and resilient scenario considering geographical constraints for ECG monitoring applications

The results in Figure 6.13 show that, for both scenarios, the energy consumption of processing increases as the level of demand increases for all number of processing servers per candidate node. This is because increasing the demand increases the number of processing servers proportionally. For all processing servers allowed at each candidate node, the energy consumed is equal for both resilience levels. This is due to fact that the same number of servers will be utilised regardless of their location, as the patients are optimally consolidated in the servers. Also, there is a slight increase in the energy consumption of processing, when the number of processing servers per candidate node increases. This is due to increasing the utilisation time of the processing servers to send the feedback and storage traffic under increasing number of processing servers per candidate node, as explained in Section 6.5.1.

6.6 Realistic parameter considerations for fall monitoring applications

This section elaborates in detail on the methodologies of determining the considered model input parameters, such as the network layout under the GPON network, the number of monitored patients in West Leeds, and the calculation of the data rate for traffic transmission of fall monitoring applications.

6.6.1 Network layout under GPON network and total number of monitored elderly patients in West Leeds, United Kingdom

In this section, video recording for fall monitoring application is considered. The patient connections and traffic routing are optimised so that the patients are served at both primary and secondary processing servers in the fog architecture, described in Section 6.2. Also, the same location of clinics and LTE base stations and the same number of patients in the clinics as described in Chapter 5, are considered. As the size of the video recording is high and due to the complexity of the MILP model (in the face of increased number of base stations), a scenario with only 7 clinics was considered. These have a total of 36 patients and 16 Long Term Evaluation (LTE) base stations were considered at the access layer, using the locations at West Leeds as case studies. The 16 LTE base stations are selected, based on the 12 nearest distances between the available base stations (BSs) and the clinics. A case study with two clusters is considered and the clinics are connected to up to three BSs in each cluster, as shown in Figure 6.14. For example, clinic 5, shown in purple, is connected to three base stations in cluster 1, and is also connected to two base stations in cluster 2.

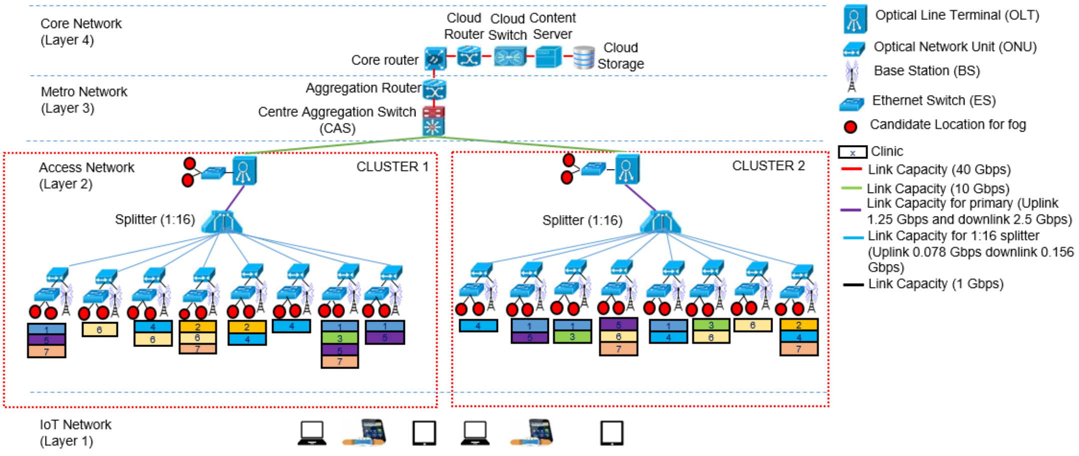


Figure 6.14: The resilient fog computing infrastructure for fall monitoring applications

Figure 6.14 shows the resilient fog computing architecture under the GPON network for fall monitoring application, while Table 6.4 presents the considered total number of elderly patients at each clinic.

Table 6.4: Number of Monitored Elderly Patients in Clinics

Clinic	Total Number of Patients
West Lodge Surgery	13
Hillfoot Surgery	2
Dr. Kw Mcgechaen & Partner	2
Pudsey Health Centre	4
Robin Lane Health	6
Dr. S M Chen & Partner	2
Manor Park Surgery	7

6.6.2 Data rate calculation for traffic transmission in the network

Just like in Chapter 5, in this section, the same 15 seconds video recording (Π) with a size of 3.36 Mbits is utilised. The patient IoT devices send their 15-second video recording to the fog servers to reconfirm the occurrence of a patient fall, based on much higher processing capabilities, compared to the IoT devices before triggering a doctor. This is to avoid a false alarm that can potentially be given to the doctors. Also, the same duration of processing and analysis for each video and the size of the analysed data (as was done in Chapter 5), which is 0.18 seconds [142] and 2.048 kbits [145], respectively are considered. The number of patients that can be served in a single processing server $Pat,$ is limited to 20% of the total number of patients from the 7 clinics, which is the lowest demand evaluated in the network. It considers different number of processing servers per candidate node $N,$ to evaluate the performance of the proposed resilient infrastructure under the fall monitoring application.

The same methodologies in Section 6.4.2, are used to determine the related data rates and the transmission time for the five approaches related to the percentage of patients. Table 6.5, shows the data rate and transmission time to transmit the video data to the processing server and to transmit the analysed data to the clinics and cloud for feedback and permanent storage, respectively, for the different number of processing servers per candidate node, N. Note that, the data rate and transmission time for each number of processing servers per candidate node is the same for all approaches. Table 6.5, also shows that the data rate and transmission time to transmit the video signal to the processing server, are the same for all the different numbers of processing servers per candidate node. This is due to the same reason as explained in Section 6.4.2.

Table 6.5: Data rate and related time for different numbers of processing servers per candidate node N for fall monitoring applications

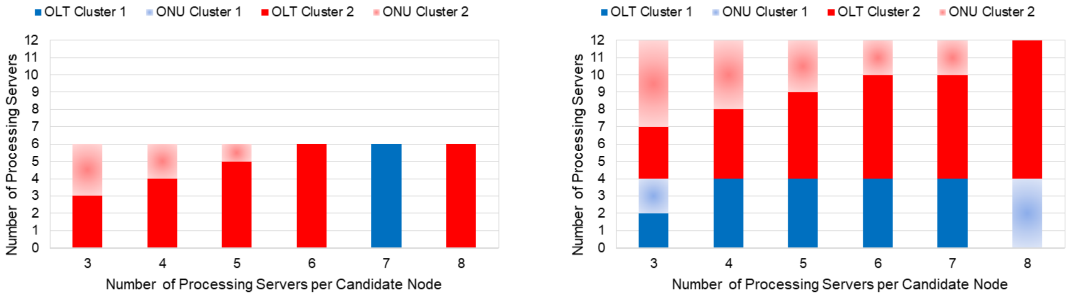
Type of Data	3 PSs	4 PSs	5 PSs	6 PSs	7 PSs	8 PSs
Data rate to transmit video data to processing server, δa (kbps)	15.120	15.120	15.120	15.120	15.120	15.120
Transmission time to the processing server, τa (s)	222.22	222.22	222.22	222.22	222.22	222.22
Data rate to transmit analysed video data to clinics, δb (kbps)	12.768	9.744	7.728	6.384	5.376	4.704
Transmission time to the clinics, τb (s)	0.16	0.21	0.265	0.321	0.381	0.435
Data rate to transmit analysed video data to cloud storage, δc (kbps)	13.020	9.765	7.812	6.510	5.580	4.882
Transmission time to the cloud storage, τc (s)	0.157	0.21	0.262	0.315	0.367	0.419

6.7 Results and analysis of the MILP model for fall monitoring applications considering server protection

In this section, the impact of increasing the level of resilience for server protection on the energy consumption of networking equipment and processing for fall monitoring applications is evaluated. The evaluations are divided into two steps, as seen in Section 6.5. The primary and secondary processing servers used to process and analyse the video signal are the same as in Section 5.4. Also, the evaluation is only performed for the highest demand level (i.e. 100% of patients).

6.7.1 Performance analysis of server protection resilient scenario without geographical constraints

In this section, the performance of the non-resilient scenario is used as a benchmark to evaluate the resilient model without geographical constraints, in terms of the energy consumption of networking equipment and processing for fall monitoring applications.



(a)

(b)

Figure 6.15: Optimal location of processing servers for (a) non-resilient scenario and (b) resilient scenario, without geographical constraints for fall monitoring applications

The results in Figure 6.15-(a) and Figure 6.15-(b), show that the number of processing servers for the resilient scenario is doubled, compared to the non-resilient scenario. This is because the non-resilient scenario does not consider secondary processing servers, as explained in Section 6.5.1. The results also show that for the non-resilient scenario, a maximum of six primary processing servers are used to serve all patients while in the resilient scenario, the total number of processing servers (i.e. primary and secondary processing servers) is twelve. The total number of processing servers in both scenarios is the same for all the different numbers of processing servers per candidate node.

Figure 6.15-(a), shows that for the non-resilient scenario, the processing servers are placed at one cluster only to reduce the total number of networking equipment (i.e. OLT) utilised. The processing servers are only placed at the OLT, when the number of processing servers allowed at each candidate node is equal to or more than the total number of processing servers required, to serve all patients. This is because the OLT is the nearest shared point to the patients. However, due to the limited capacity of the OLT to support all of the traffic, the ONUs at the same cluster are selected by the MILP to place the remaining processing servers.

Meanwhile, the resilient scenario without geographical constraints for server protection has resulted in utilising the BSs, ONUs and OLTs in both clusters. This is because, for the resilient scenario, each patient required double resources from the base stations to send the video data to both primary and secondary processing servers. Therefore, due to the high data rate to send the video signal and the restricted number of available resources of the base stations in a single cluster, the processing servers are placed at both clusters under the resilient scenario. Also, increasing the number of processing servers per candidate node, has resulted in placing the processing servers at both OLTs. This is because the OLTs are the nearest shared point, hence reducing the number of candidate nodes utilised to place the processing servers and the total traffic traversing the network.

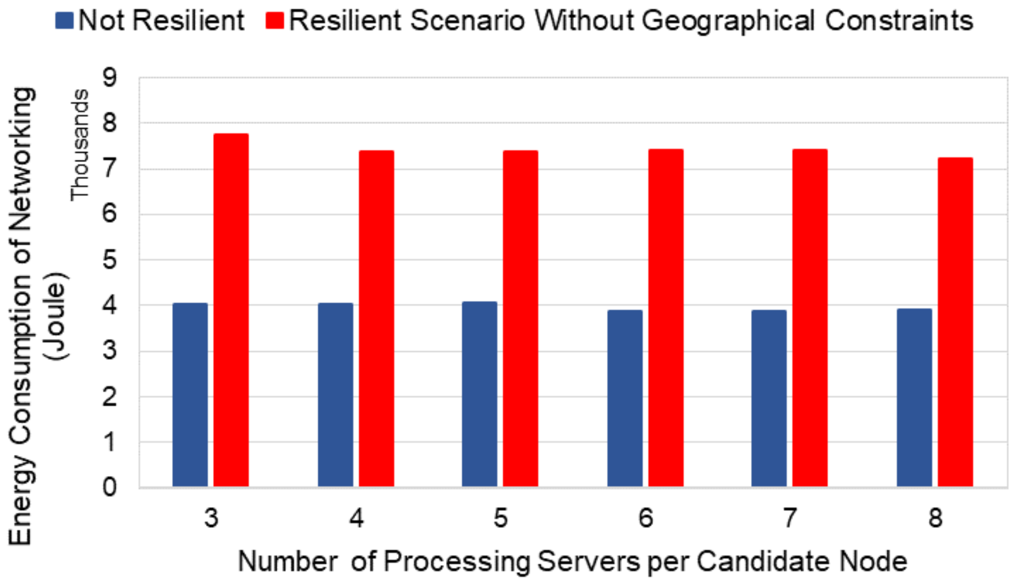


Figure 6.16: Energy consumption of networking equipment for non-resilient and resilient scenario, without geographical constraints for fall monitoring applications

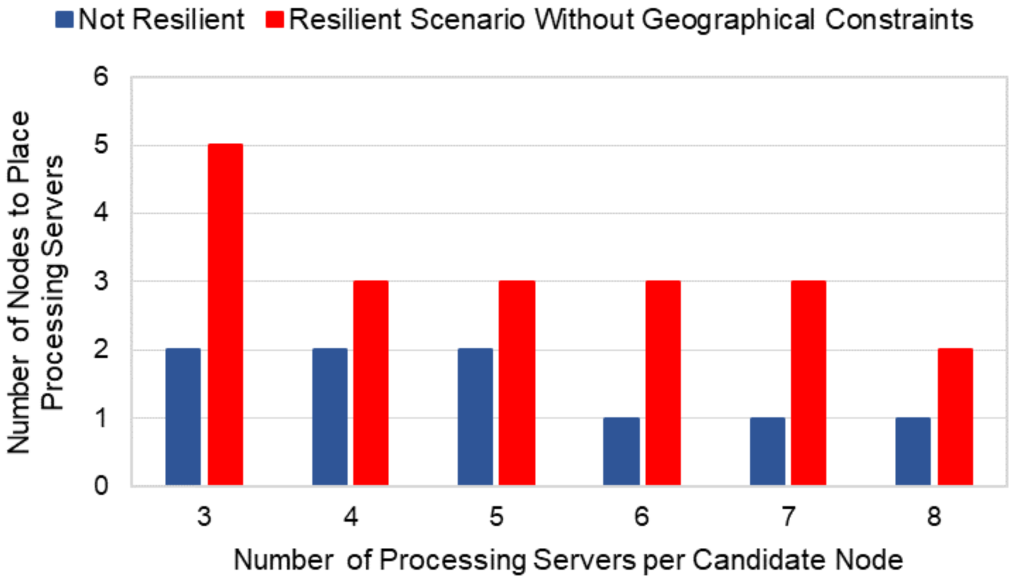


Figure 6.17: Total number of candidate nodes used to place the processing servers for non-resilient and resilient scenario, without geographical constraints for fall monitoring applications

The results in Figure 6.16, show that the energy consumed in the resilient scenario is always higher than that of the non-resilient scenario. This is due to the high traffic traversing the networking equipment in the resilient scenario, as explained in Section 6.5.1. Figure 6.16 also shows that when the number of processing servers per candidate node increases from three to eight, the energy consumption of networking equipment in the resilient scenario and the non-resilient scenario reduces significantly. The reduction in the energy consumption of the networking equipment, is due to the optimised placement of the processing servers and the reduction in the number of candidate nodes utilised to place the processing servers, as explained in Section 6.5.1. The reduction in the number of candidate nodes utilised to place the processing servers in both scenarios is as depicted in Figure 6.17.

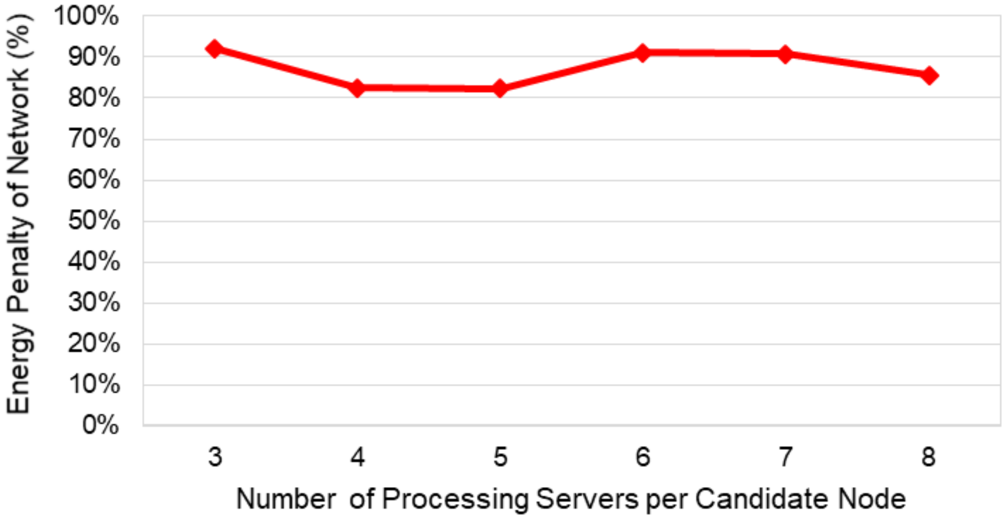
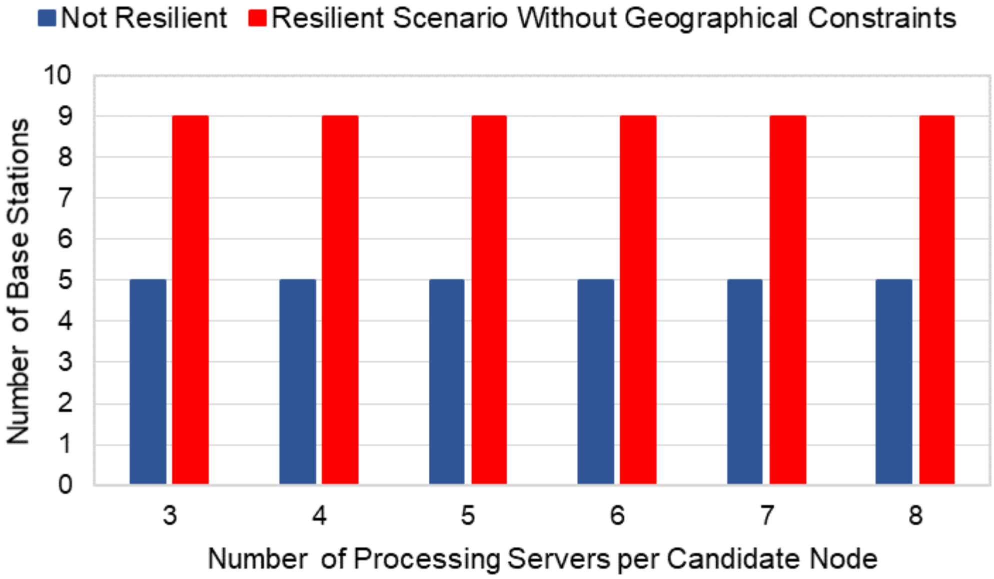
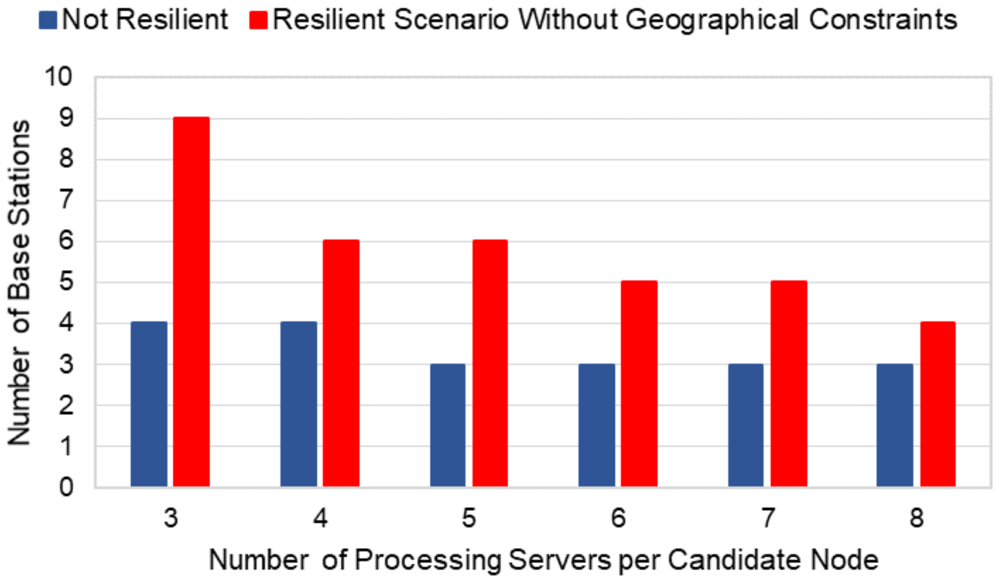


Figure 6.18: Energy penalty of networking equipment for resilient scenario without geographical constraints, compared to the non-resilient scenario for fall monitoring applications.



(a)



(b)

Figure 6.19: Number of base stations used to send (a) the raw video signal for processing and (b) analysed video signal for feedback, for the non-resilient scenario and resilient scenario, without geographical constraints under different number of processing servers per candidate node for fall monitoring applications

The results in Figure 6.18, show that the energy penalty of the network due to considering resilience for server protection is high (i.e. more than 80%). This is because the total traffic traversing the network in the resilient scenario is doubled that of the non-resilient scenario. Therefore, more networking equipment (i.e. base stations, Ethernet switches), are utilised in the resilient scenario to serve the high traffic. This is shown in Figure 6.19-(a), Figure 6.19-(b) and Figure 6.17; where the number of base stations utilised to send the raw video signal to the processing servers for processing and analysis, the number of the base stations used to send the analysed health data traffic to the clinics and the number of candidate nodes used to place the processing servers in the resilient scenario without geographical constraints, respectively are higher than that of the non-resilient scenario. However, it is worth noting that the energy consumed by the base stations to send the feedback traffic, does not give significant increases to the total energy consumption of networking equipment, due to its low utilisation time.

The results also show that increasing the number of processing servers per candidate node, can either decrease or increase the energy penalty, due to the increasing level of resilience. The increase in energy penalty with the increase in the number of processing servers allowed per candidate node, is due to the reduction in the number of candidate nodes needed to place the processing servers under the non-resilient scenario. On the other hand, the energy penalty decreases with the increase in the number of processing servers allowed per candidate node. This is due to the reduction in the number of candidate nodes needed to place the processing servers under the resilient scenario. The reduction in the number of candidate nodes used to place the processing servers with the increasing number of processing servers allowed per candidate node for both scenarios, is illustrated in Figure 6.17.

Meanwhile, the energy consumption of processing for the resilient scenario, is always higher than that of the non-resilient scenario, and this energy increased in both scenarios when the number of processing servers per candidate node increased. This is due to the same reason as discussed in Section 6.5.1. The results of energy consumption of processing for non-resilient scenario and resilient scenario without geographical constraints, for fall monitoring applications can be found in Appendix 3.

6.7.2 Performance analysis of resilient scenario with geographical constraints

In this section, the performance of the resilient scenario without geographical constraints, is used as a benchmark. The more resilient scenario with geographical constraints is then evaluated and compared to the benchmark in terms of its energy consumption of networking equipment and processing for fall monitoring applications.

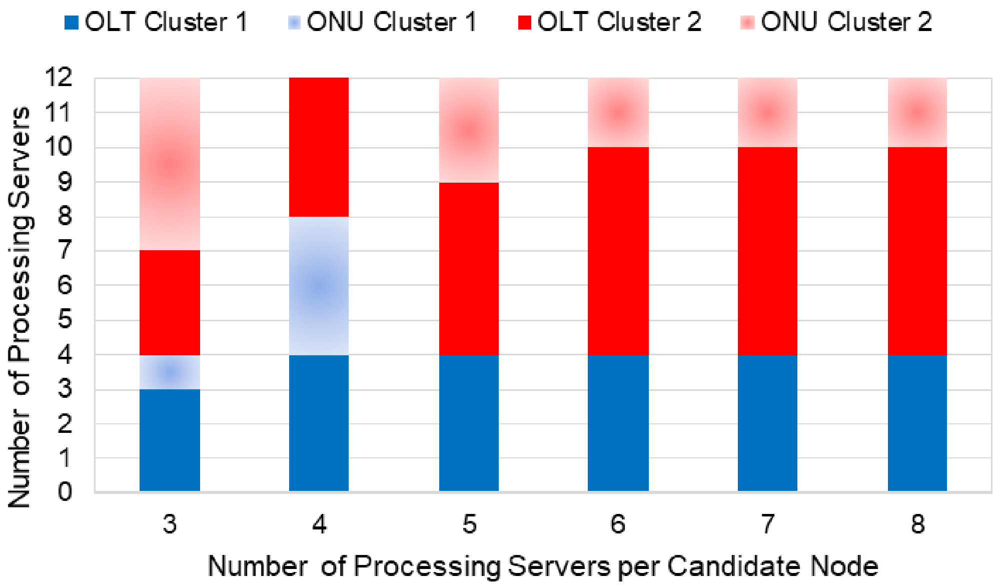


Figure 6.20: Optimal location of processing servers for the resilient scenario with geographical constraints, for fall monitoring applications

As in Section 6.5.2, the results in Figure 6.20 show that the processing servers are placed in both clusters, due to the limited resources of the base stations in a single cluster to serve all patients. The results show that when three processing servers are allowed at each candidate node, the OLT and

ONU of both clusters are used to place the increased number of processing servers. This is because, the OLTs do not have enough capacity to place all the processing servers; therefore, the remaining patients are served by servers at the ONUs.

However, the results show that by increasing the number of processing servers per candidate node, this has resulted in placing more processing servers at the OLT. The reason for doing this, is to reduce the amount of data traversing the OLTs, so that data can be sent to the processing servers at the ONUs. Also, increasing the number of processing servers per candidate node has resulted in placing the processing servers at optimal locations. This is shown in Figure 6.20, where the same numbers and locations are used to place the processing servers, when the number processing servers per candidate node is equal to or more than six, respectively.

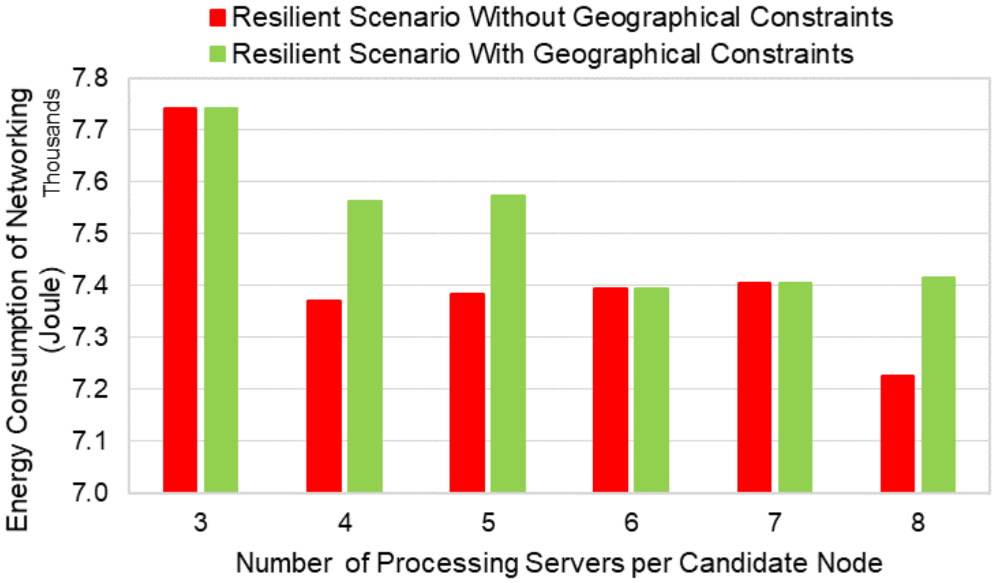


Figure 6.21: Energy consumption of networking equipment for scenario with geographical constraints; and scenario without geographical constraints for fall monitoring applications.

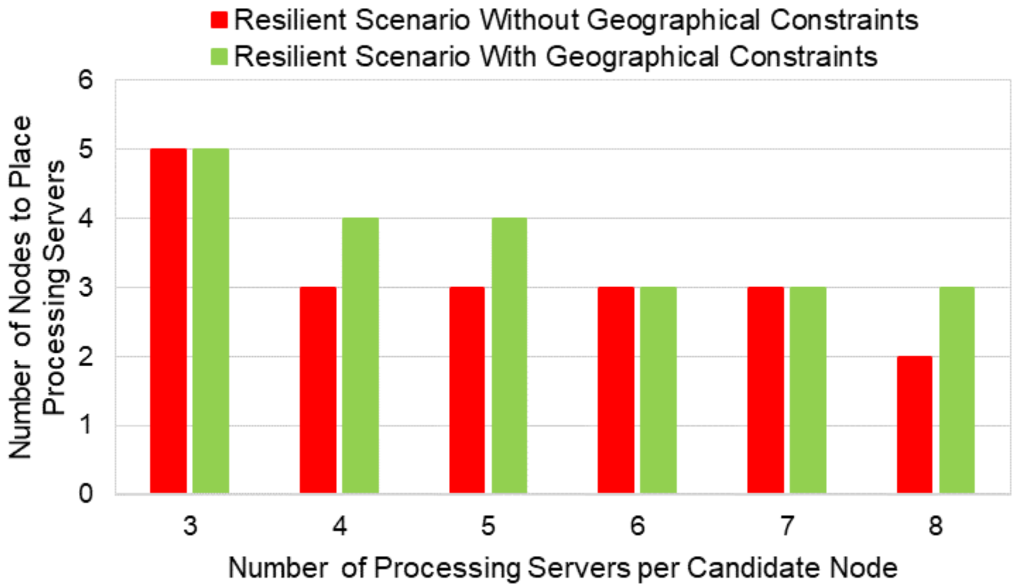


Figure 6.22: Total number of candidate nodes used to place the processing servers for the resilient scenario, with geographical constraints and scenario without geographical constraints for fall monitoring applications.

The results in Figure 6.21, show that increasing the level of resilience can either maintain or increase the energy consumption of networking equipment. This depends on the number of processing servers that can be connected at each candidate node. The same level of energy of networking equipment consumed in both scenarios, is due to the same number of candidate nodes being utilised to place the processing servers, as shown in Figure 6.22 and the same number of base stations used to serve the patients, to send the video signal to the processing servers and to send the analysed video signal to the clinics, which can be found in Appendix 4.

On the other hand, the increasing energy consumption of networking equipment due to the increasing level of resilience; is because of the high number of candidate nodes utilised to place the processing servers in the more resilient scenario, as the primary and secondary processing servers are not allowed to be placed at the same location. This is shown in Figure 6.22, where the number of candidate nodes used to place the processing servers for the more resilient scenario, is always equal to or higher than the resilient scenario, without a geographical constraint. The high amount of energy consumed for the more resilient scenario, is the penalty for having a higher level of resilience.

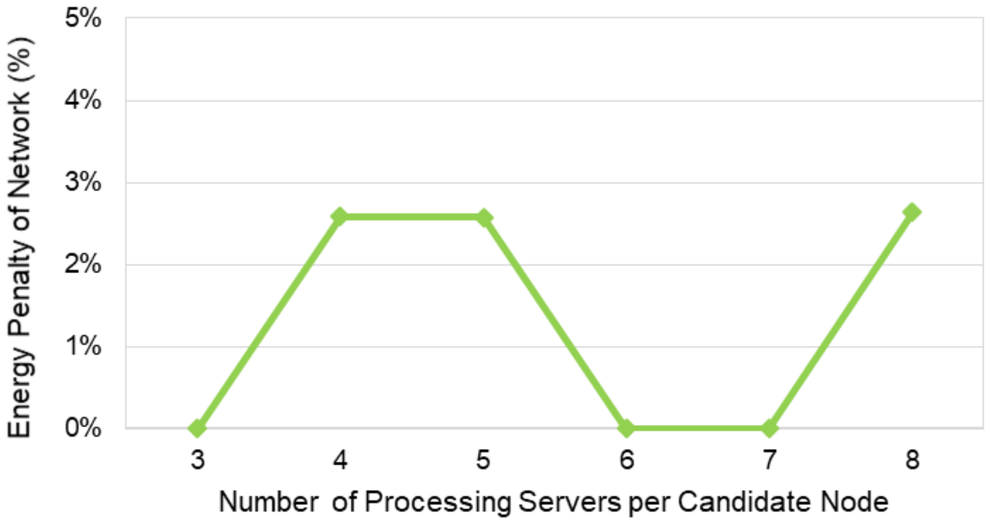


Figure 6.23: The percentage energy penalties of networking equipment for the scenario with geographical constraints, compared to the scenario without geographical constraints for fall monitoring applications

The results in Figure 6.23, also show that when three, six and seven processing servers are available at each candidate node, increasing the level of resilience does not incur an energy penalty. This is mainly due to the same number of candidate nodes being utilised to place the processing servers as shown in Figure 6.22 and the same number of base stations being utilised to send the video signal to the processing server; then to send the analysed video to the clinics (the results can be found in Appendix 4).

The results also show that increasing the number of processing servers allowed at each candidate node, can increase or decrease the energy penalty, due to the increasing level of resilience. The increase in the energy penalty with the increase in the number of processing servers allowed per candidate node, is due to the reduction in the number of candidate nodes needed to place the processing servers under the resilient scenario, without geographical constraints. This shows that increasing the number of processing servers per candidate node, can only profit the scenario without geographical constraints. However, the increase in the energy penalty is less than 3%, as shown in Figure 6.23. On the other hand, the energy penalty decreases with the increase in the number of processing servers per candidate node and that is due to the reduced number of candidate nodes needed to place the processing servers under the more resilient scenario.

Meanwhile, the energy consumption of processing in both scenarios is equal, and this energy increased as the number of processing servers per candidate node increased. This is due to the same reason as explained Section 6.5.2. The results of energy consumption of processing for the resilient scenario without geographical constraints and the resilient scenario, considering geographical constraints for fall monitoring applications can be found in Appendix 5.

6.8 Resilient infrastructure with server and network protection

In this section, we design an energy-efficient fog computing infrastructure for health monitoring applications, resilient against server and network failures. We consider the same 1+1 protection scheme as in previous resilient scenarios considering server protection, where two servers, a primary server and a secondary server, are used to serve the health monitoring application concurrently. Since geographic constraints do not add a lot of power, therefore in this section only server protection with geographic constraints is considered. Two types of network protection are considered in addition to server protection offering a scenario with higher levels of resilience. In this scenario, the primary and secondary processing servers are not allowed to be placed at the same node, and the links and nodes used to transmit the data to and from primary and secondary processing servers are disjoint, as node and link failures in the network are not improbable. We consider the disjoint links and nodes to be only at the access layer, as the processing servers can only be placed at the access layer. A Mixed Integer Linear Programming (MILP) model is used to optimise the number and placement of the primary and secondary processing servers so that the energy consumption of both networking equipment and processing are minimised. As in Section 6.5, AMPL software with CPLEX 12.8 solver running on high-performance computing (HPC) cluster with a 12 core CPU and 64 GB RAM was used as the platform for solving the MILP models. We investigate the performance of the proposed resilience architecture in terms of energy consumption of both networking equipment and processing under (Electrocardiogram) ECG monitoring and video fall monitoring applications, separately. For each application, the patients will send the required data (i.e. 30 seconds ECG signal for ECG monitoring applications and 15 seconds video recording for the fall monitoring application) to the primary and secondary processing servers for processing, analysis and decision-making. In addition, we observe the energy penalty of increasing the level of resilience in the network. Also, we developed a heuristic model for the considered resilience scenario for real-time implementation.

6.8.1 Mathematical model for energy-efficient fog computing considering server and network protection

In this section, a mathematical model for the resilient scenario that considers geographical constraints for server protection; and link and node disjoint resilience for network protection are introduced. The Mixed Integer Linear Programming (MILP) model with the objective of minimising the total energy consumption of both networking equipment and processing of the resilience scenario is developed. Here the primary and secondary processing servers are not allowed to be placed at the same node, and the links and nodes used to relay the traffic to and from both primary and secondary processing servers are disjoint. Beyond the OLT and heading to the cloud, the network is not protected, since the server that did the processing has a copy of the data to be stored and can retain it until the network beyond the OLT recovers. Note that the disjoint links and nodes are considered to be only at the access layer. The same parameters, variables, constraints and objective functions in Section 6.3.2 are utilised and an additional set and variables, as shown in Table 6.6 (also can be found in Appendix 1), are introduced to optimise the number and locations of the primary and secondary processing servers considering the geographical constraints and link and node disjoint resilience, so that the energy consumption of both networking equipment and processing are minimised.

Table 6.6: Additional set and variables in the MILP model

Set	
ND	Set of BSs, ONUs and OLTs (access layer)
Variables	
Pa_{sd}	Raw health data traffic from clinic s to primary processing servers at destination node d (bps), $s \in CL, d \in FN$
Pb_{sd}	Raw health data traffic from source node s to secondary processing servers at destination node d (bps), $s \in CL, d \in FN$
Pa_{ij}^{sd}	Raw health data traffic from source node s to primary processing servers at destination node d that traverses the link between nodes i and j (bps), $s \in CL, d \in FN,\ i, j \in N$
Pb_{ij}^{sd}	Raw health data traffic from source node s to secondary processing servers at destination node d that traverses the link between nodes i and j (bps), $s \in CL, d \in FN,\ i, j \in N$
Fa_{sd}	Analysed health data feedback traffic from primary processing servers at source node s to clinic at node d (bps), $s \in FN, d \in CL$
Fb_{sd}	Analysed health data feedback traffic from secondary processing servers at source node s to clinic at node d (bps), $s \in FN, d \in CL$
Fa_{ij}^{sd}	Analysed health data feedback traffic from primary processing servers at source node s to clinic at node d that traverses the link between nodes i and j (bps), $s \in FN, d \in CL, i, j \in N$
Fb_{ij}^{sd}	Analysed health data feedback traffic from secondary processing servers at source node s to clinic at node d that traverses the link between nodes i and j (bps), $s \in FN, d \in CL, i, j \in N$

Sa_{sd}	Analysed health data storage traffic from primary processing servers at source node s to cloud storage at node d (bps), $s \in FN, d \in CST$
Sb_{sd}	Analysed health data storage traffic from secondary processing servers at source node s to cloud storage at node d (bps), $s \in FN, d \in CST$
Sa_{ij}^{sd}	Analysed health data storage traffic from primary processing servers at source node s to cloud storage at node d that traverses the link between nodes i and j (bps), $s \in FN, d \in CST, i, j \in N$
Sb_{ij}^{sd}	Analysed health data storage traffic from secondary processing servers at source node s to cloud storage at node d that traverses the link between nodes i and j (bps), $s \in FN, d \in CST, i, j \in N$
La_{ij}	$La_{ij} = 1$, if the incoming and/or outgoing traffic of primary processing servers traverses the link between nodes i and j otherwise $La_{ij} = 0$
Lb_{ij}	$Lb_{ij} = 1$, if the incoming and/or outgoing traffic of secondary processing servers traverses the link between nodes i and j otherwise $Lb_{ij} = 0$
ρa_i	$\rho a_i = 1$, if the incoming and/or outgoing traffic of primary processing servers traverse node i, otherwise $\rho a_i = 0$
ρb_i	$\rho b_i = 1$, if the incoming and/or outgoing traffic of secondary processing servers traverses node i, otherwise $\rho b_i = 0$

In addition to constraints presented in Section 6.3.2, the following new constraints are considered:

1) Traffic from clinics to fog.

$$Pa_{sd} = \omega a_{sd}\ \delta a \quad ;\ s \in CL, d \in FN \qquad (6\text{-}20)$$

$$Pb_{sd} = \omega b_{sd}\ \delta a \quad ;\ s \in CL, d \in FN \qquad (6\text{-}21)$$

Constraints (6-20), (6-21) calculate the raw health data traffic from clinic s to the primary and secondary processing servers located at node d, respectively. This is based on the association of patients from clinic to processing servers (i.e. ωa_{sd} and ωb_{sd}), as well as the data rate provisioned for each patient, $\delta a,$ to perform the transmission.

2) Traffic from fog to clinics.

$$Fa_{sd} = \omega a_{sd}\ \delta b \quad ;\ s \in FN, d \in CL \qquad (6\text{-}22)$$

$$Fb_{sd} = \omega b_{sd}\ \delta b \quad ;\ s \in FN, d \in CL \qquad (6\text{-}23)$$

Constraints (6-22), (6-23) calculate the analysed health data feedback traffic from primary and secondary processing servers located at node s to the clinic d, respectively. This is based on the association of patients from clinic to processing servers (i.e. ωa_{sd} and ωb_{sd}), as well as the data rate provisioned for each patient, $\delta b,$ to perform the transmission.

3) Traffic from fog to cloud storage.

$$Sa_{sd} = \sum_{i \in CL} \omega a_{is} \ \delta c \ \delta_{sd} \quad ; \ s \in FN, d \in CST \tag{6-24}$$

$$Sb_{sd} = \sum_{i \in CL} \omega b_{is} \ \delta c \ \delta_{sd} \quad ; \ s \in FN, d \in CST \tag{6-25}$$

Constraints (6-24), (6-25) calculate the analysed health data storage traffic from primary and secondary processing servers located at node s to cloud storage, d respectively. This is based on the association of patients from clinic to processing servers (i.e. ωa_{is} and ωb_{is}), as well as the data rate provisioned for each patient, $\delta c,$ to perform the transmission. Note that, in this work, there is only one cloud storage d, therefore the δ_{sd} is a parameter equal that is equal to 1.

4) Flow conservation in the network.

$$\sum_{j \in Nm[i]: i \neq j} Pa_{ij}^{sd} - \sum_{j \in Nm[i]: i \neq j} Pa_{ji}^{sd} = \begin{cases} Pa_{sd} \ if \ i = s \\ -Pa_{sd} \ if \ i = d \\ 0 \ otherwise \end{cases} \tag{6-26}$$

$$s \in CL, d \in FN, i \in N$$

$$\sum_{j\in Nm[i]:i\neq j} Pb_{ij}^{sd} - \sum_{j\in Nm[i]:i\neq j} Pb_{ji}^{sd} = \begin{cases} Pb_{sd}\, if\ i = s \\ -Pb_{sd}\, if\ i = d \\ 0\ otherwise \end{cases} \tag{6-27}$$

$$s \in CL, d \in FN, i \in N$$

$$\sum_{j\in Nm[i]:i\neq j} Fa_{ij}^{sd} - \sum_{j\in Nm[i]:i\neq j} Fa_{ji}^{sd} = \begin{cases} Fa_{sd}\, if\ i = s \\ -Fa_{sd}\, if\ i = d \\ 0\ otherwise \end{cases} \tag{6-28}$$

$$s \in FN, d \in CL, i \in N$$

$$\sum_{j\in Nm[i]:i\neq j} Fb_{ij}^{sd} - \sum_{j\in Nm[i]:i\neq j} Fb_{ji}^{sd} = \begin{cases} Fb_{sd}\, if\ i = s \\ -Fb_{sd}\, if\ i = d \\ 0\ otherwise \end{cases} \tag{6-29}$$

$$s \in FN, d \in CL, i \in N$$

$$\sum_{j\in Nm[i]:i\neq j} Sa_{ij}^{sd} - \sum_{j\in Nm[i]:i\neq j} Sa_{ji}^{sd} = \begin{cases} Sa_{sd}\, if\ i = s \\ -Sa_{sd}\, if\ i = d \\ 0\ otherwise \end{cases} \tag{6-30}$$

$$s \in FN, d \in CST, i \in N$$

$$\sum_{j\in Nm[i]:i\neq j} Sb_{ij}^{sd} - \sum_{j\in Nm[i]:i\neq j} Sb_{ji}^{sd} = \begin{cases} Sb_{sd}\, if\ i = s \\ -Sb_{sd}\, if\ i = d \\ 0\ otherwise \end{cases} \tag{6-31}$$

$$s \in FN, d \in CST, i \in N$$

Constraints (6-26) – (6-31) ensure that the total incoming traffic is equivalent to the total outgoing traffic for all nodes in the network, except for source and destination nodes for processing, feedback and storage tasks, respectively.

5) Link used to transmit raw and analysed health data traffic.

$$\sum_{s\in CL}\sum_{d\in FN} Pa_{ij}^{sd} + \sum_{s\in FN}\sum_{d\in CL} Fa_{ij}^{sd} + \sum_{s\in FN}\sum_{d\in CST} Sa_{ij}^{sd} \geq La_{ij} \tag{6-32}$$

$$i \in N, j \in Nm[i]$$

$$\sum_{s\in CL}\sum_{d\in FN} Pa_{ij}^{sd} + \sum_{s\in FN}\sum_{d\in CL} Fa_{ij}^{sd} + \sum_{s\in FN}\sum_{d\in CST} Sa_{ij}^{sd} \leq M\ La_{ij} \tag{6-33}$$

$$i \in N, j \in Nm[i]$$

$$\sum_{s\in CL}\sum_{d\in FN} Pb_{ij}^{sd} + \sum_{s\in FN}\sum_{d\in CL} Fb_{ij}^{sd} + \sum_{s\in FN}\sum_{d\in CST} Sb_{ij}^{sd} \geq Lb_{ij} \tag{6-34}$$

$$i \in N, j \in Nm[i]$$

$$\sum_{s\in CL}\sum_{d\in FN} Pb_{ij}^{sd} + \sum_{s\in FN}\sum_{d\in CL} Fb_{ij}^{sd} + \sum_{s\in FN}\sum_{d\in CST} Sb_{ij}^{sd} \leq M\ Lb_{ij} \tag{6-35}$$

$$i \in N, j \in Nm[i]$$

Constraints (6-32), (6-33) ensure that $La_{ij} = 1$ if the incoming and/or outgoing traffic of primary processing servers traverses the link between nodes i and j, otherwise the value is zero. Meanwhile, Constraints (6-34), (6-35) ensure that the $Lb_i = 1$, if the incoming and/or outgoing traffic of secondary processing servers traverses the link between nodes i and j, otherwise the value is zero.

6) Disjoint links constraint.

$$La_{ij} + Lb_{ij} \leq 1 \tag{6-36}$$

$$i \in ND, j \in ND$$

$$La_{ij} + Lb_{ji} \leq 1 \tag{6-37}$$

$$i \in ND, j \in ND$$

$$La_{ji} + Lb_{ij} \leq 1 \tag{6-38}$$

$$i \in ND, j \in ND$$

Constraints (6-36) – (6- 38) ensure that the incoming and/or outgoing traffic of the primary and secondary processing servers traverse different links.

7) Disjoint nodes constraint.

$$\sum_{j \in Nm[i]: i \neq j} La_{ij} \geq \rho a_i \tag{6-39}$$

$$i \in ND$$

$$\sum_{j \in Nm[i]: i \neq j} La_{ij} \leq M\, \rho a_i \tag{6-40}$$

$$i \in ND$$

$$\sum_{j \in Nm[i]: i \neq j} Lb_{ij} \geq \rho b_i \tag{6-41}$$

$$i \in ND$$

$$\sum_{j \in Nm[i]: i \neq j} Lb_{ij} \leq M\ \rho b_i \tag{6-42}$$

$$i \in ND$$

$$\rho a_i + \rho b_i \leq 1 \tag{6-43}$$

$$i \in ND$$

Constraints (6-39), (6-40) and constraints (6-41), (6-42) determine the nodes that are used to relay the incoming and/or outgoing traffic of the primary processing server and secondary processing servers, respectively. Meanwhile, constraint (6-43) ensures that the nodes used to relay the incoming and/or outgoing traffic of primary and secondary processing servers are different.

6.8.2 Results and analysis of the MILP model for ECG monitoring applications considering geographical constraint server protection and link and node disjoint

In this section, we evaluate the impact of increasing the level of resilience for server and network protection on the energy consumption of networking equipment and processing. The evaluation is performed by comparing the resilient scenario with the geographical constraint (i.e. benchmark) with the more resilient scenario considering additional link and node disjoint routing for ECG monitoring applications. As in Section 6.5, the locations of the processing servers of the evaluated scenarios, the energy consumption of the networking equipment and processing of the scenarios and the energy penalty of networking equipment due to the increasing level of resilience are analysed.

The same resilient architecture and the parameter inputs used in Section 6.4 for the ECG monitoring applications are utilised to evaluate the resilience scenario.

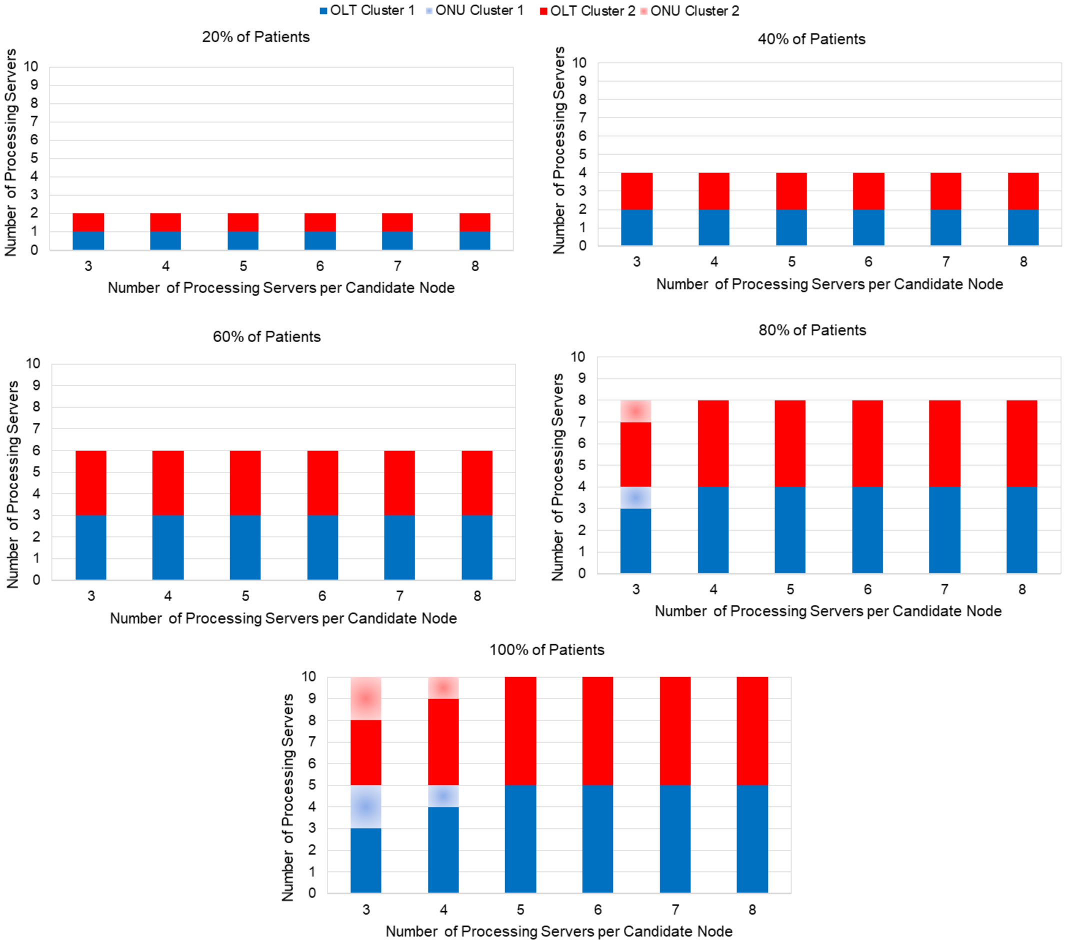


Figure 6.24: Optimal location of processing servers for the resilient scenario considering the geographical constraint for server protection and link and node disjoint resilience for network protection for ECG monitoring applications

The results in Figure 6.24 show that the processing servers are only placed at the (optical line terminals) OLTs in both clusters when the number of processing servers allowed at each candidate node is equal to or greater than the total number of primary or secondary processing servers required in the network. This is for two reasons. The first is to reduce the number of candidate nodes (i.e. Ethernet switches) used to place the processing servers, as the OLTs are the nearest shared point to the patients. The second is because each cluster is used to place the same set of processing servers. For instance, cluster 1 is used to place only primary processing servers, while cluster 2 is used to place only secondary processing servers. Therefore, when the number of processing servers allowed at each candidate node is less than the number of primary and secondary processing servers required, the (optical network units) ONUs in both clusters are utilised to place the remaining processing servers under increasing demands.

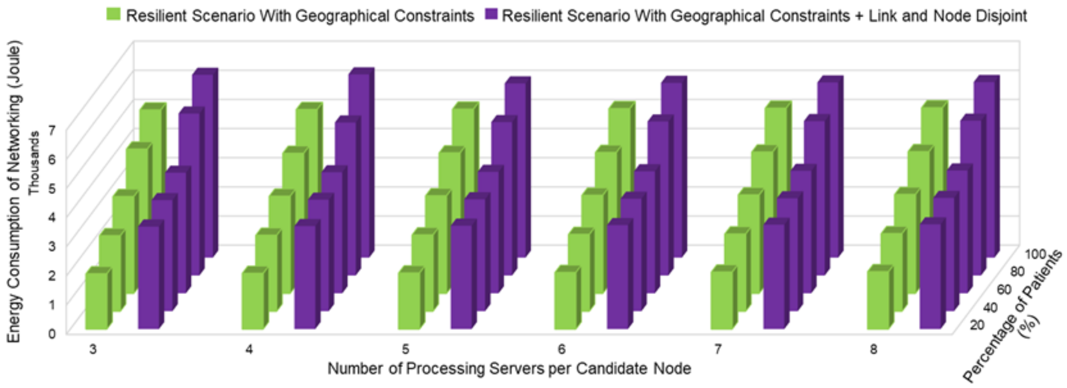
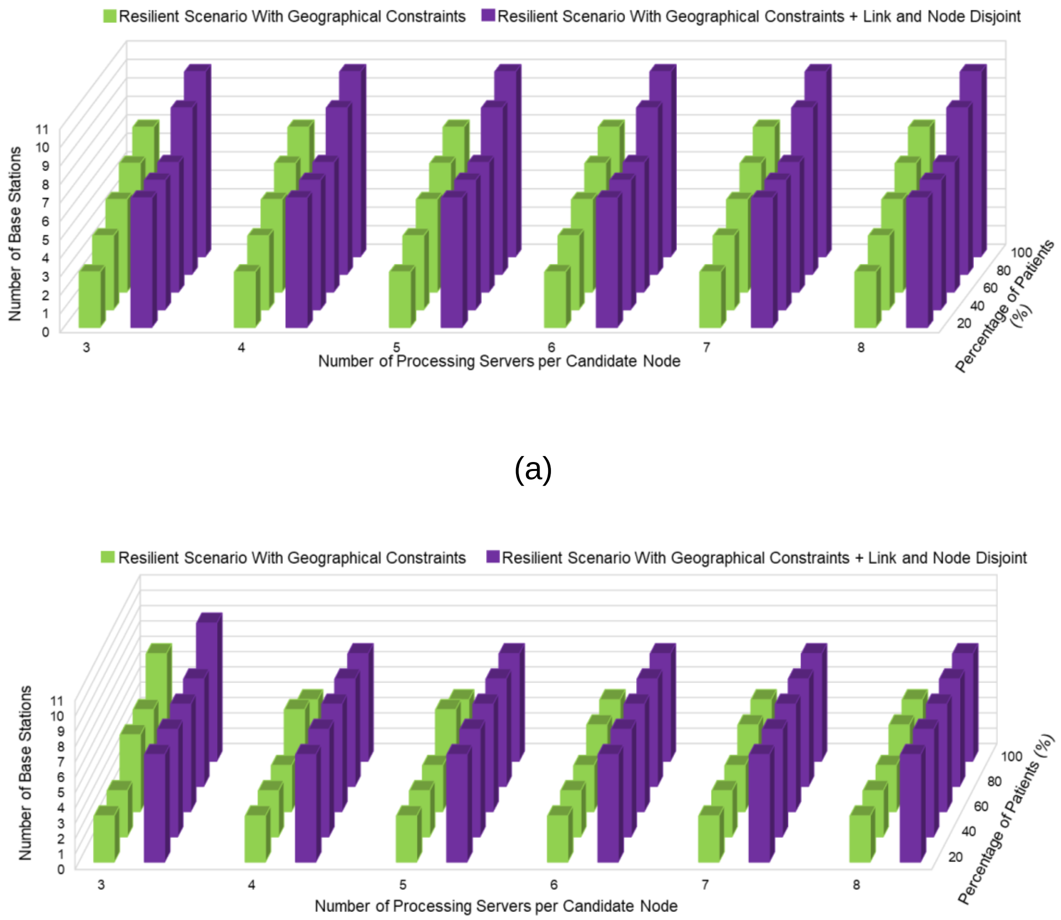


Figure 6.25: Energy consumption of networking equipment for the resilient scenario considering the geographical constraints and the resilient scenario with geographical constraints and link and node disjoint resilience for ECG monitoring applications



(a)

(b)

Figure 6.26: Number of base stations used to send the (a) raw ECG signal for processing and (b) analysed ECG signal for feedback, for the resilient scenario considering the geographical constraints; and the resilient scenario considering geographical constraints and link and node disjoint resilience under different percentages of patients and number of processing servers per candidate node

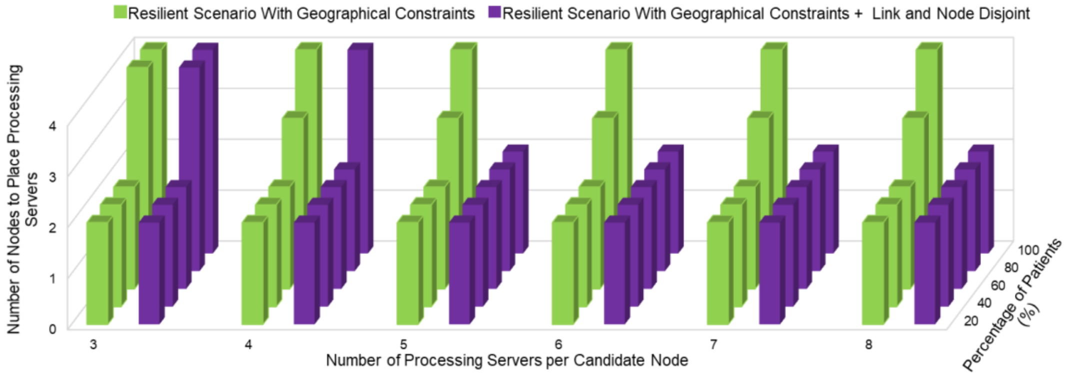


Figure 6.27: Number of candidate nodes used to place the processing servers for the resilient scenario considering the geographical constraints and the resilient scenario with geographical constraints and link and node disjoint resilience for ECG monitoring applications

The results in Figure 6.25 show that the energy consumption of networking equipment for both scenarios increases as the demand increases for all the different numbers of processing servers per candidate node considered. This is due to the increasing amount of traffic in the network, hence increasing the total number of networking equipment utilised in the network.

The results also show that, for all levels of demands and number of processing servers per candidate node, the energy consumption of networking equipment for the more resilient scenario is always higher than the resilient scenario that only considers geographical constraints. This is due to the high number of base stations utilised in the more resilient scenario, as shown in Figure 6.26-(a) and Figure 6.26-(b). Note that, each base station is connected to only one OLT in the network. Therefore, considering disjoint links and nodes for network protection has increased the number of base stations without maximising the utilisation of their resources to send the processing traffic to both primary and secondary processing servers.

It is worth noting that the number of candidate nodes used to place the processing servers at demand levels of 80% and 100% in the more resilient scenario is lower than the resilient scenario with geographical constraints when the number of processing servers per candidate node is equal to or more than four and five, respectively, as shown in Figure 6.27. However, as the energy consumed by a single base station is approximately 1.5x higher than the energy consumed by a single node (i.e. Ethernet switch) to place the processing servers, therefore there is an energy penalty with the link and node disjoint resilience scenario.

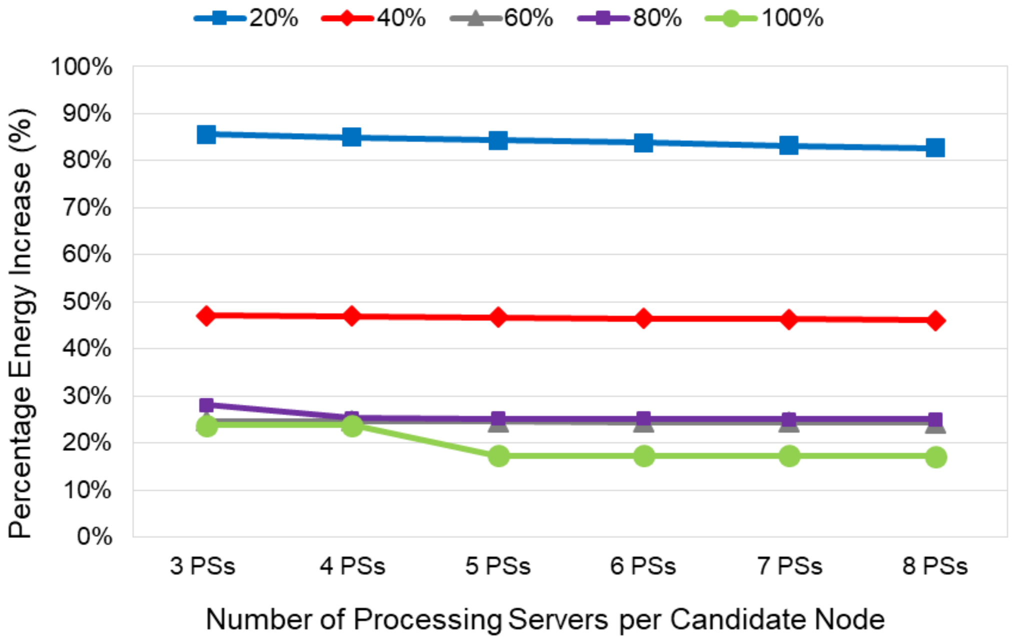


Figure 6.28: Percentage energy penalty of networking equipment for the resilient scenario considering the geographical constraints and link and node disjoint resilience compared to the resilient scenario considering the geographical constraints for ECG monitoring applications

The results in Figure 6.28 show that the energy penalty with the link and node disjoint resilience scenario decreases as the demand level increases from 20% to 60% and 80% to 100%. This is because the total number of base stations utilised in the resilient scenario, that only consider the geographical constraint, increases with the increases in demand in the network, as shown in Figure 6.26-(a) and Figure 6.26-(b). This increases the energy consumption of networking equipment for the resilient scenario that only considers the geographical constraint as the demand levels increase. However, at a demand level of 60%, the energy penalty is lower than at a demand level of 80%. This is because, at a demand level of 80%, the number of base stations used for the more resilient scenario start to increase, hence increasing the energy consumption of networking equipment of the more resilient scenario.

Figure 6.28 also shows that, at demand levels of 80% and 100%, increasing the number of processing servers per candidate node to 4 and 5, respectively, decreases the energy penalty. This is because, the number of candidate nodes (i.e. Ethernet switches) used to place the processing servers for the more resilient scenario reduces while the same number of candidate nodes are used for the resilient scenario that only considers geographical constraints as shown in Figure 6.27.

Meanwhile, increasing the level of resilience does not increase the energy consumption of processing, and this energy increased as the number of processing servers per candidate node increased. This is due to the same reason as explained Section 6.5.2. The results of energy consumption of processing for the resilient scenario considering the geographical constraints; and the energy consumption of the resilient scenario with geographical constraints and link and node disjoint resilience for ECG monitoring applications can be found in Appendix 6.

6.8.3 Results and analysis of the MILP model for fall monitoring applications considering geographical constraint server protection and link and node disjoint

In this section, the impact of increasing the level of resilience through network protection on the energy consumption of networking equipment and processing for fall monitoring applications is evaluated. The performance of the resilient scenario with geographical constraints is used as a benchmark to evaluate the power consumption implications (in terms of the energy consumption of networking equipment and processing) of increasing the level of resilience. The additional resilience is achieved when additional link and node disjoint resilience is considered for fall monitoring applications. The evaluations are performed, as in Section 6.8.2. The same resilient architecture and the parameter inputs used in Section 6.6 for the fall monitoring applications are utilised to evaluate the resilience scenario. As in Section 6.7, the evaluation is performed for the highest level of demand (i.e. 100% of patients).

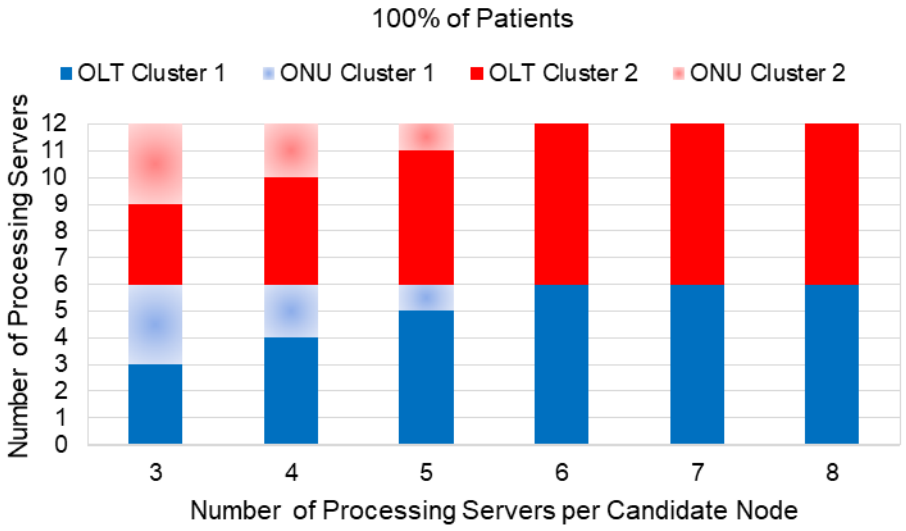


Figure 6.29: Optimal location of processing servers for the resilient scenario considering the geographical constraint for server protection and link and node disjoint resilience for network protection for fall monitoring applications

The results in Figure 6.29 show the same patterns as in Section 6.8.2 when considering a resilient scenario with geographical constraints and link and node disjoint resilience. Here the processing servers are only placed at the OLTs when the number of processing servers allowed at each candidate node is equal to or greater than the total number of primary or secondary processing servers required in the network. This is for the same two reasons as explained in Section 6.8.2. Figure 6.29 also shows that, when the number of processing servers allowed at each candidate node is equal to or less than five, the OLTs of both clusters host the maximum number of processing servers they can serve. Therefore, the ONUs in both clusters are used to place the remaining processing servers that cannot be allocated at the OLTs due to the limited number of processing servers per candidate node.

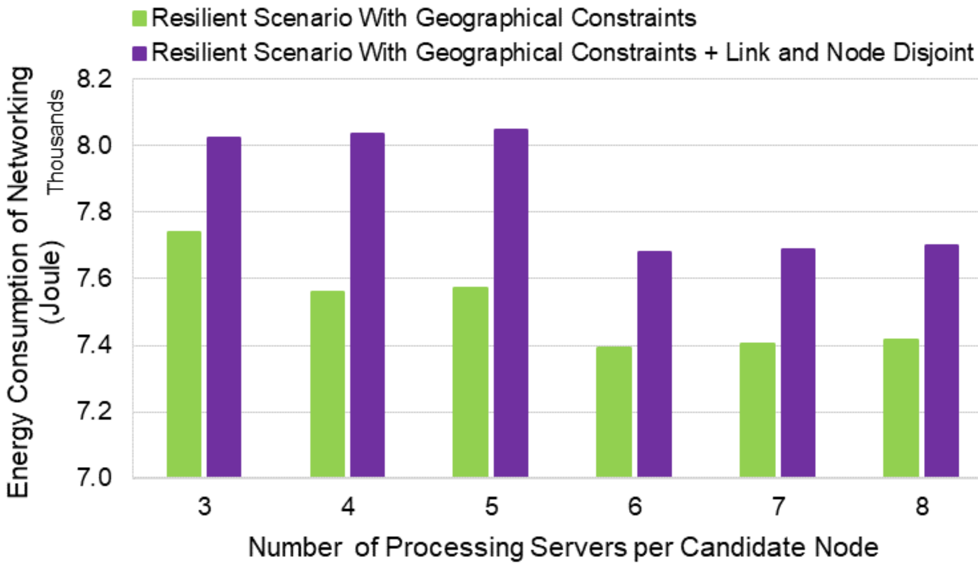


Figure 6.30: Energy consumption of networking equipment for the resilient scenario considering the geographical constraints; and the resilient scenario with geographical constraints; and link and node disjoint resilience for fall monitoring applications

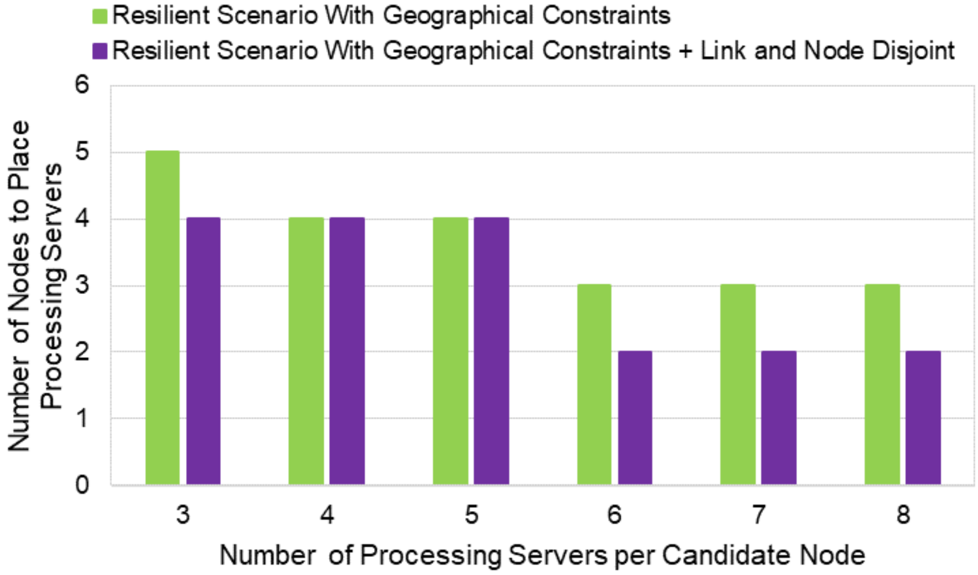
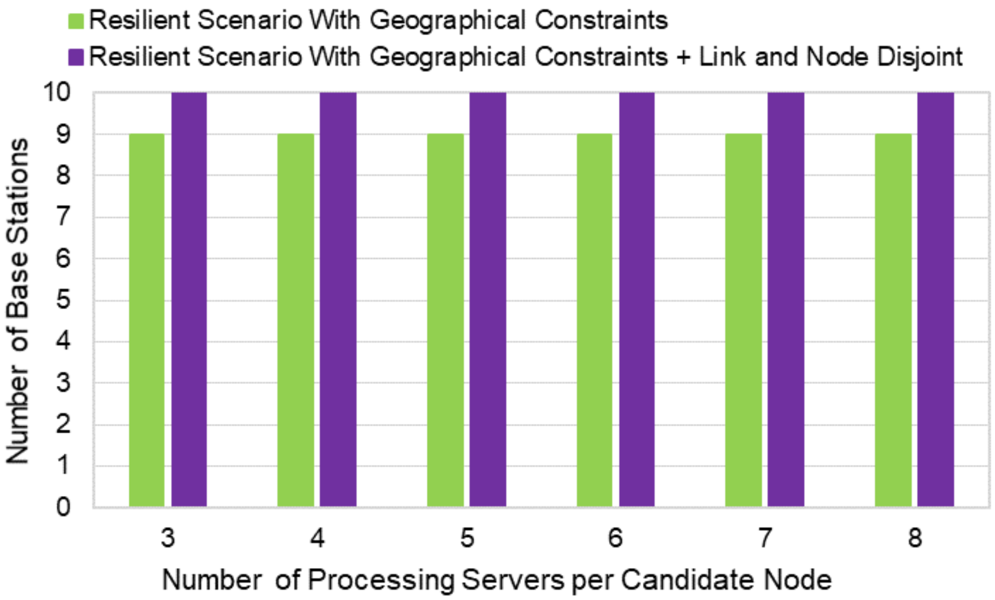
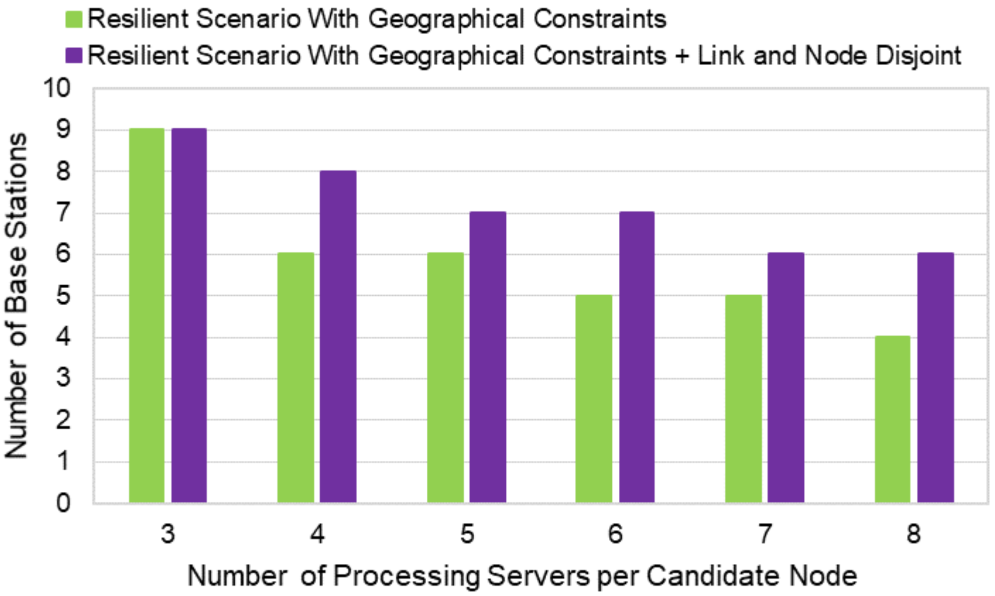


Figure 6.31: Total number of candidate nodes used to place processing servers for the resilient scenario considering the geographical constraints; and the resilient scenario with geographical constraints and link and node disjoint resilience for fall monitoring applications



(a)



(b)

Figure 6.32: Number of base stations used to send (a) the raw video signal for processing and (b) the analysed video signal for feedback, for the resilient scenario considering the geographical constraints; and the resilient scenario with geographical constraints and link and node disjoint resilience for different numbers of processing servers per candidate node

The results in Figure 6.30 show that, for all the different number of processing servers per candidate node, the energy consumption of networking equipment in the more resilient scenario is always higher than the scenario that only considers geographical constraints. This is due to the higher total number of utilised base stations in the more resilient scenario to send both raw health data traffic and analysed health data feedback traffic, compared to the resilient scenario that only considered the geographical constraint, as illustrated in Figure 6.32-(a) and Figure 6.32-(b). The reason for the increase in the number of the base stations in the more resilient scenario is as explained in Section 6.8.2. Also, as shown in Figure 6.30, increasing the number of processing servers per candidate node can also reduce the energy consumption of networking equipment in both scenarios, when the number of candidate nodes utilised to place the processing servers is reduced as illustrated in Figure 6.31.

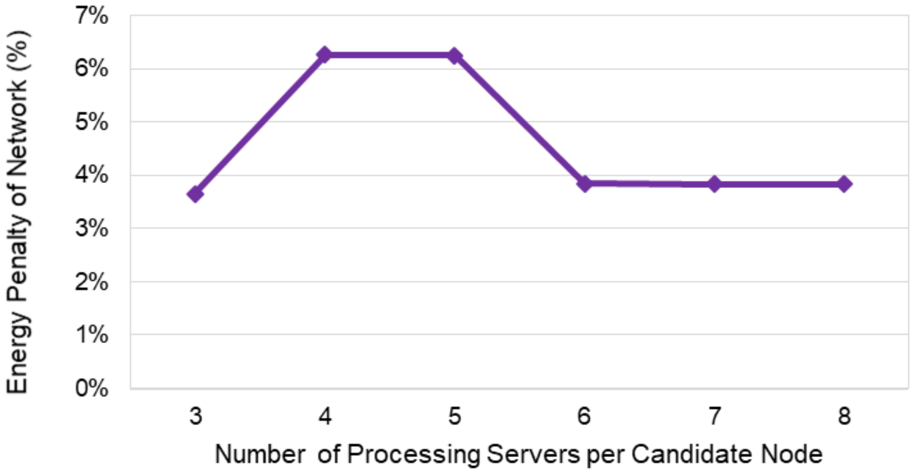


Figure 6.33: Percentage energy penalty of networking equipment for the resilient scenario considering the geographical constraints and link and node disjoint resilience compared to the resilient scenario considering the geographical constraints for fall monitoring applications

The results in Figure 6.33 show that increasing the level of resilience to consider link and node disjoint resilience for network protection compared to the resilient scenario that only considers geographical constraints has resulted in an energy penalty. This is mainly due to the high number of base stations and the number of candidate nodes required to place the processing servers utilised in the more resilient scenario. The results also show that the energy penalty can increase and decrease with the increasing number of processing servers per candidate node. The increase in energy penalty with the increase in the number of processing servers that can be served at each candidate node is due to the reduction in the number of candidate nodes needed to place the processing servers under the resilient scenario that only considers geographical constraints. On the other hand, the energy penalty decreasing with the increased number of processing servers that can be served at each candidate node is due to the reduction in the number of candidate nodes to place the processing servers under the more resilient scenario. The reduction in the number of candidate nodes used to place the processing server with the increasing number of processing servers allowed per candidate node for both scenarios is illustrated in Figure 6.31.

Meanwhile, the same energy for processing is consumed in both resilient scenarios, and this energy increased as the number of processing servers per candidate node increased. This is due to the same reasons as explained in Section 6.5.2. The results of energy consumption of processing for the resilient scenario considering the geographical constraints and the resilient scenario considering the geographical constraints and link and node disjoint resilience for fall monitoring applications can found in Appendix 7.

6.9 Energy optimised resilient infrastructure with fog computing heuristic

The Energy optimised resilient infrastructure fog computing without geographical constraints (EORIWG) heuristic, the Energy optimised resilient infrastructure fog computing with geographical constraints (EORIG) heuristic and Energy optimised resilient infrastructure fog computing with geographical constraints and link and node disjoint (EORIGN) heuristic were developed to validate the MILP operation and to deliver a real-time solution of the resilient scenario without geographical constraints, with geographical constraints, and both with geographical constraints and link and node disjoint, respectively. The heuristics are developed based on the insights from the results obtained from the MILP models. The operations of the EORIWG, EORIG and EORIGN heuristics are discussed based on the given flow charts as follows.

6.9.1 Flow of EORIWG heuristic

The heuristic determines the BSs to be used to serve patients to send raw health data and receive feedback data and the nodes to place primary and secondary processing servers at the access network so that the energy consumption of both networking and processing are minimised. Figure 6.34 shows the flow chart of the EORIGW heuristic.

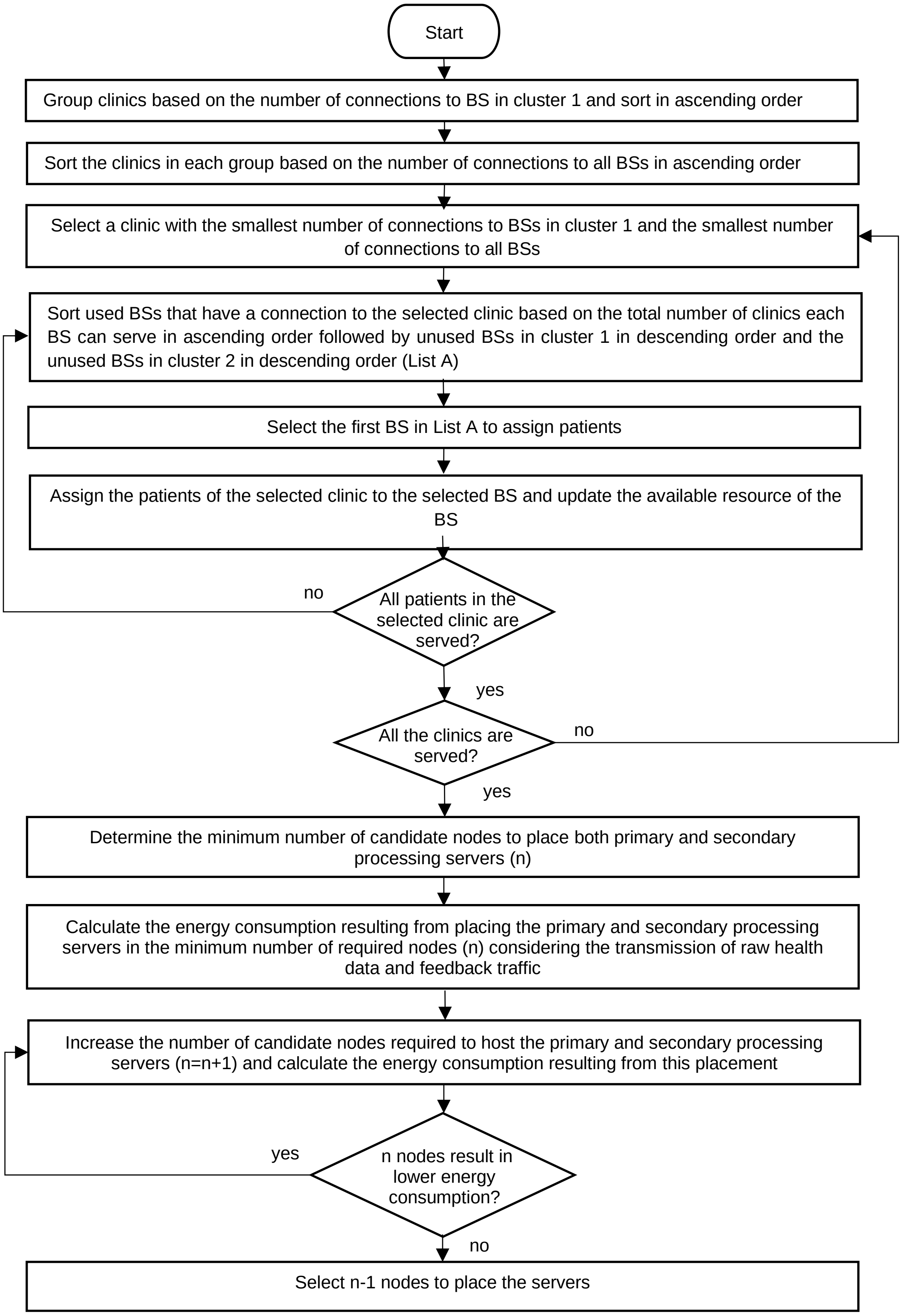

Start
Group clinics based on the number of connections to BS in cluster 1 and sort in ascending order
Sort the clinics in each group based on the number of connections to all BSs in ascending order
Select a clinic with the smallest number of connections to BSs in cluster 1 and the smallest number of connections to all BSs
Sort used BSs that have a connection to the selected clinic based on the total number of clinics each BS can serve in ascending order followed by unused BSs in cluster 1 in descending order and the unused BSs in cluster 2 in descending order (List A)
Select the first BS in List A to assign patients
Assign the patients of the selected clinic to the selected BS and update the available resource of the BS
All patients in the selected clinic are served?
no
yes
All the clinics are served?
no
yes
Determine the minimum number of candidate nodes to place both primary and secondary processing servers (n)
Calculate the energy consumption resulting from placing the primary and secondary processing servers in the minimum number of required nodes (n) considering the transmission of raw health data and feedback traffic
Increase the number of candidate nodes required to host the primary and secondary processing servers (n=n+1) and calculate the energy consumption resulting from this placement
n nodes result in lower energy consumption?
yes
no
Select n-1 nodes to place the servers

Figure 6.34: Flow chart for EORIGW heuristic

The heuristic begins by grouping the clinics based on the number of BSs in cluster 1 it can connect to and sorts the groups in ascending order. For each group, the clinics are sorted based on the total number of BSs in both clusters the clinic can connect to in ascending order. The heuristic assigns first the clinic with the smallest number of connections to the BSs in cluster 1 and the smallest number of connections to all BSs in both clusters, to the BSs to help in reducing the utilisation of OLTs. Also, it ensures that all clinics are assigned to BSs.

The assignment of clinic patients to a BS is as follows: The heuristic sorts the BSs that have a connection to the clinic under consideration starting with BSs previously used by the healthcare application that has available resources. These BSs are sorted in ascending order based on the total number of clinics the BS can serve followed by the unused BSs in cluster 1 in descending order and followed by the unused BSs in cluster 2 also in descending order. Sorting the activated BSs in ascending order is used to reduce the number of utilised BSs while the descending order of unused BSs in cluster 1 followed by the unused BSs in cluster 2 is used to ensure that options are left open until late in the allocation process while minimising the utilisation of the OLTs. Then, the patients of the clinic under consideration are consolidated to the minimum number of BSs to reduce the number of BSs used by the healthcare application. Note that, for each patient, the heuristic assigned double resources to clinics so that they send their health data to both primary and secondary processing servers.

The heuristic then determines the number of primary and secondary processing servers required to serve the patients and the nodes hosting them.

The candidate nodes to host the servers are the ONUs connected to the BSs selected to serve the patients and the OLTs. Considering the minimum number of candidate nodes required to host both primary and secondary servers to serve all the patients (which is based on the maximum number of servers a node can host), the heuristic finds the combination of candidate nodes to host the primary and secondary processing servers that result in minimum energy consumption. Limiting the number of candidate nodes to place the primary and secondary processing servers reduces the utilisation of the Ethernet switches to serve the processing servers.

The energy consumption that results from hosting both primary and secondary processing servers at a combination of candidate nodes is calculated by routing the traffic (raw health data traffic) from BSs (starting with the BS serving the largest number of patients) to the nearest node with available processing capacity of the combination of candidate nodes under consideration based on minimum hop routing.

Also, BSs to send feedback traffic from combination of candidate nodes to clinics are selected using the same approach used to select BSs to send raw health data. Note that BSs different from those used to send raw health data are used to send feedback traffic. The difference is for the same reason as explained for the EOFC heuristic in Section 4.4.1.

The combination of nodes hosting servers considering the minimum number of candidate nodes required to host primary and secondary servers to serve all the patients that result in minimum energy consumption is selected.

The heuristic increases the number of candidate nodes to host servers and repeats the above process. The decision to select the location to host the primary and secondary processing servers with the increasing number of candidate node is the same as explained in Section 4.4.1 for EOFC heuristic.

6.9.2 Flow of the EORIG heuristic

The EORIG heuristic determines the BSs to serve patients so as to send raw health data and receive feedback data and the nodes to place primary and secondary processing servers at the access network so that the energy consumption of both networking and processing is minimised and the primary and secondary servers are node disjoint (geographical constraints). Below is the list of the changes made for the EORIG heuristic compared to EORIGW heuristic:

1. The number of candidate nodes to place processing servers is based on the total number of candidate nodes to place primary and secondary processing servers in disjoint nodes.
2. Assigning patients from BSs to the primary processing servers is done first and the nodes used to place the primary processing servers are removed from the combination of nodes before assigning the same patients from the BSs to the secondary processing servers.

6.9.3 Flow of the EORIGN heuristic

The heuristic determines the BSs to be used to serve the patients so as to send the raw health data and receive feedback data. It also determines the nodes to be used to place the primary and the secondary processing servers at the access network so that the energy consumption of both networking and processing are minimised. Figure 6.35 shows the flow chart of the EORIGN heuristic.

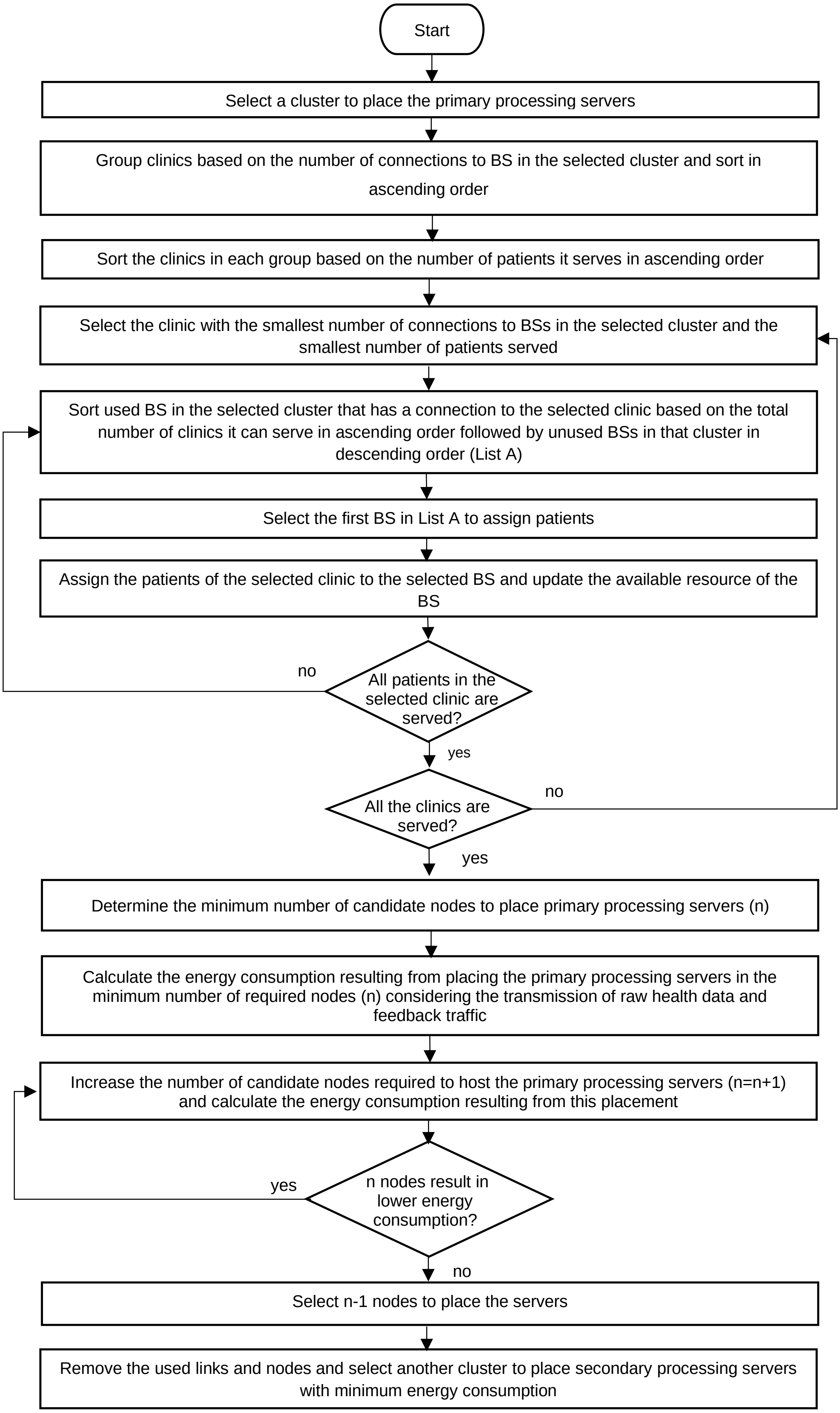

Start
Select a cluster to place the primary processing servers
Group clinics based on the number of connections to BS in the selected cluster and sort in ascending order
Sort the clinics in each group based on the number of patients it serves in ascending order
Select the clinic with the smallest number of connections to BSs in the selected cluster and the smallest number of patients served
Sort used BS in the selected cluster that has a connection to the selected clinic based on the total number of clinics it can serve in ascending order followed by unused BSs in that cluster in descending order (List A)
Select the first BS in List A to assign patients
Assign the patients of the selected clinic to the selected BS and update the available resource of the BS
no
All patients in the selected clinic are served?
yes
no
All the clinics are served?
yes
Determine the minimum number of candidate nodes to place primary processing servers (n)
Calculate the energy consumption resulting from placing the primary processing servers in the minimum number of required nodes (n) considering the transmission of raw health data and feedback traffic
Increase the number of candidate nodes required to host the primary processing servers (n=n+1) and calculate the energy consumption resulting from this placement
yes
n nodes result in lower energy consumption?
no
Select n-1 nodes to place the servers
Remove the used links and nodes and select another cluster to place secondary processing servers with minimum energy consumption

Figure 6.35: Flow chart of EORIGN heuristic

In the EORIGN heuristic, the selection of the locations to host the primary servers and secondary servers are done separately to ensure that the traffic to the primary server and the traffic to the secondary servers are routed separately (link and node disjoint). In this process, the heuristic begins by selecting a cluster to assign the patients in the clinics to the primary processing server. Then the heuristic groups the clinics based on the number of BSs in the selected cluster it can connect to and sorts the groups in ascending order. For each group, the clinics are sorted based on the total number of patients it serves in ascending order. The heuristic assigns first the clinic with the smallest number of connections to the BSs in the selected cluster and the smallest number of patients it serves to the BSs to ensure each clinic can be served by at least one BS and to help in packing the BSs (packing is optimum when equipment have high idle power consumption).

The assignment of clinic patients to a BS is as follows: The heuristic sorts the BSs in the selected cluster that has a connection to the clinic under consideration starting with BSs previously used by the healthcare application that has available resources. These BSs are sorted in ascending order based on the total number of clinics the BS can serve followed by the unused BSs in the selected cluster in descending order. The ascending order of activated BSs is used to reduce the number of utilised BS while, the descending order of unused BS in the selected cluster is used to ensure that options are left open until late in the allocation process. Then, the patients of the clinic under consideration are consolidated in the minimum number of BSs to reduce the number of BSs used by the healthcare application.

The heuristic then determines the number of primary processing servers required to serve the patients and the nodes hosting them. The candidate nodes to be used to host the servers are the ONUs connected to the BSs selected to serve the patients and the OLT of the selected cluster. Considering the minimum number of candidate nodes required to host servers to serve all the patients (which is based on the maximum number of servers a node can host), the heuristic finds the combination of candidate nodes to host the primary processing servers that results in minimum energy consumption. Limiting the number of candidate nodes used to place the primary processing servers is for the same reason as explained in Section 4.4.1 which is to reduce the number of Ethernet switches used to serve the processing servers.

The energy consumption that results from hosting servers at a combination of candidate nodes in the selected cluster is calculated as explained for EOFC heuristic in Section 4.4.1. The BSs to be used to send feedback traffic from combination of candidate nodes to clinics are selected using the same approach used to select BSs to send raw health data. Note that different BSs are used to send raw health data and to send feedback traffic for the same reason as explained in EOFC heuristic in Section 4.4.1.

The combination of nodes hosting servers considering the minimum number of candidate nodes required to host primary processing servers to serve all the patients that result in minimum energy consumption is selected.

As in EOFC heuristic in Section 4.4.1, the heuristic increases the number of candidate nodes used to host servers. The energy consumption resulting from using this combination of nodes is calculated and compared to the energy consumption resulting from the combination of nodes hosting servers considering the minimum number of candidate nodes required to host servers. If the latter is lower, the heuristic examines placing servers in more candidate nodes. If the former is lower, the minimum number of candidate nodes required to host servers is selected to place servers.

Next, the heuristic removes the links and nodes used to send the traffic to or from primary processing servers and selects another cluster to assign patients in the clinics to the secondary processing servers. Different clusters are used to host the primary and secondary processing servers, which is due to the link and node disjoint resilience mandated for network protection. The same process is used to allocate patients to the BSs to send raw health data and to receive analysed health data feedback. It is also used for the selection of locations to host the primary processing servers and to determine the optimal location to host the secondary processing server.

6.10 Results and analysis of the heuristic models

In this section, we evaluate the performance of the developed heuristics for server protection, the EORIWG heuristic and EORIG heuristics, and heuristic for server and network protection, EORIGN heuristic, compared to the MILP results in term of the energy consumption of networking equipment and processing. The evaluations are performed for both ECG monitoring applications and fall monitoring applications considering all levels of demand and 100% of demand level, respectively. As in the previous chapters, the heuristics are running on a normal PC with 3.2 GHz CPU and 16 GB RAM.

6.10.1 EORIWG heuristic results

Figure 6.36-(a) for ECG monitoring applications shows that the total energy consumption of EORIWG heuristic is equal to that of the MILP model when the demand levels are 20% and 40% for all number of processing servers per candidate node. This is due to the ability to use the minimum number of primary and secondary processing servers and the number of candidate nodes to place the processing servers that are built into the EORIWG heuristic while assigning the patients from clinics to the processing servers.

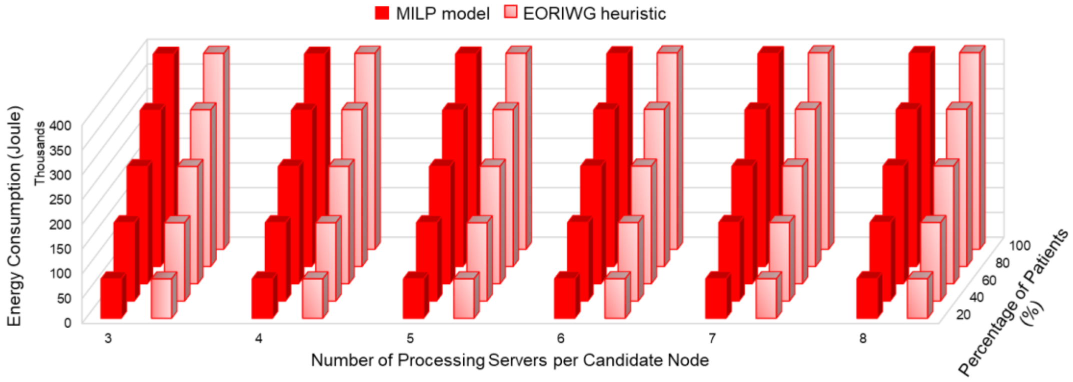
Figure 6.36-(a) also shows that the total energy consumption of the EORIWG heuristic is higher than the MILP model with an average of 0.17%, 0.42% and 0.44%, at demand levels of 60%, 80% and 100%, respectively as shown in Table 6.7. The higher energy consumed in the EORIWG heuristic is because at demand levels of 60% and 100%, increasing the patients has resulted in utilising more base stations to send the raw ECG data to the processing servers as shown in Figure 6.37-(a). In the EORIGW heuristic, all base stations in cluster 1 are utilised, and due to the different connections of each clinic to the base stations, the utilisation of the resources in the selected base stations are not maximised. Therefore, the base stations in cluster 2 are also used to serve the patients from the remaining clinics.

Also, at demand levels of 80% and 100%, the number of base stations utilised in the EORIGW heuristic to send the feedback traffic is higher than in the MILP model, as shown in Figure 6.37-(a), hence more networking equipment energy is consumed in the EORIWG heuristic compared to the MILP model. Note that, increasing the number of base stations to send the processing traffic results in more impact on the energy of networking equipment compared to the growing number of base stations used to send the feedback traffic. Also note that, in EORIWG heuristic, the number of candidate nodes used to place the processing servers is equal to the minimum required candidate nodes to place both primary and secondary processing servers. Therefore, due to the restricted number of candidate nodes to place the processing servers, the centre aggregation switch (CAS) is activated in the EORIWG heuristic to send the ECG signal to the processing servers located at different clusters when the demand levels increase to or more than 60%. The utilisation of the CAS has increased the energy consumption of networking equipment in the EORIWG heuristic.

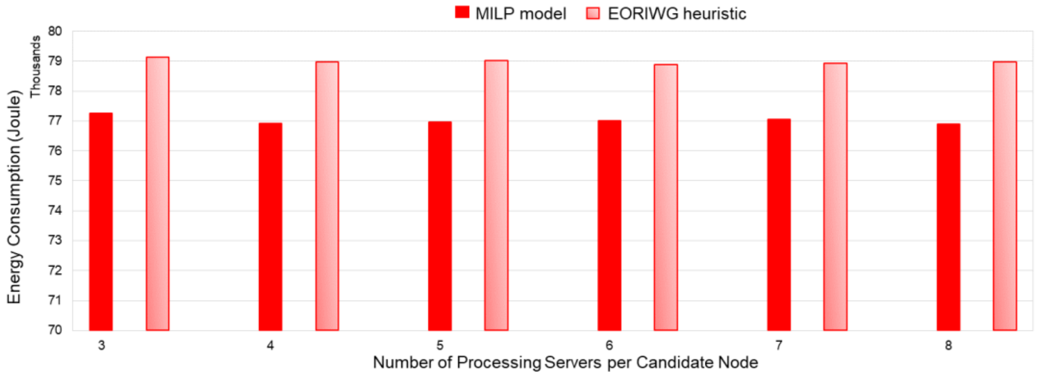
Meanwhile, Figure 6.36-(b) for fall monitoring applications, shows that, at a demand level of 100%, the total energy consumption in EORIWG heuristic is higher than the MILP model for all the different numbers of processing servers per candidate node. The average energy increase in the EORIWG heuristic compared to MILP model is 2.56% as shown in Table 6.7. The increased energy consumed in the EORIWG heuristic is for the same reasons as explained for the ECG monitoring application when the demand level is equal to or more than 60%. Note that, the high energy consumed in the EORIWG heuristic is also due to the increasing number of base stations needed to send the analysed health data feedback traffic shown in Figure 6.37-(b).

We also evaluated the computational time needed to run the EORIGW heuristic and the MILP model. The results show that, for 100% of patients and three processing servers per candidate node, the EORIGW heuristic running on a normal PC with 3.2 GHz CPU and 16 GB RAM took 86 sec and 35.3 sec to finish for ECG and fall monitoring applications, respectively. Meanwhile, the

MILP model running on high-performance computing (HPC) cluster with a 12 core CPU and 64 GB RAM took a longer time than the heuristics with 4 hours and 42 minutes for ECG monitoring applications and 7 hours and 47 minutes for fall monitoring applications.



(a)

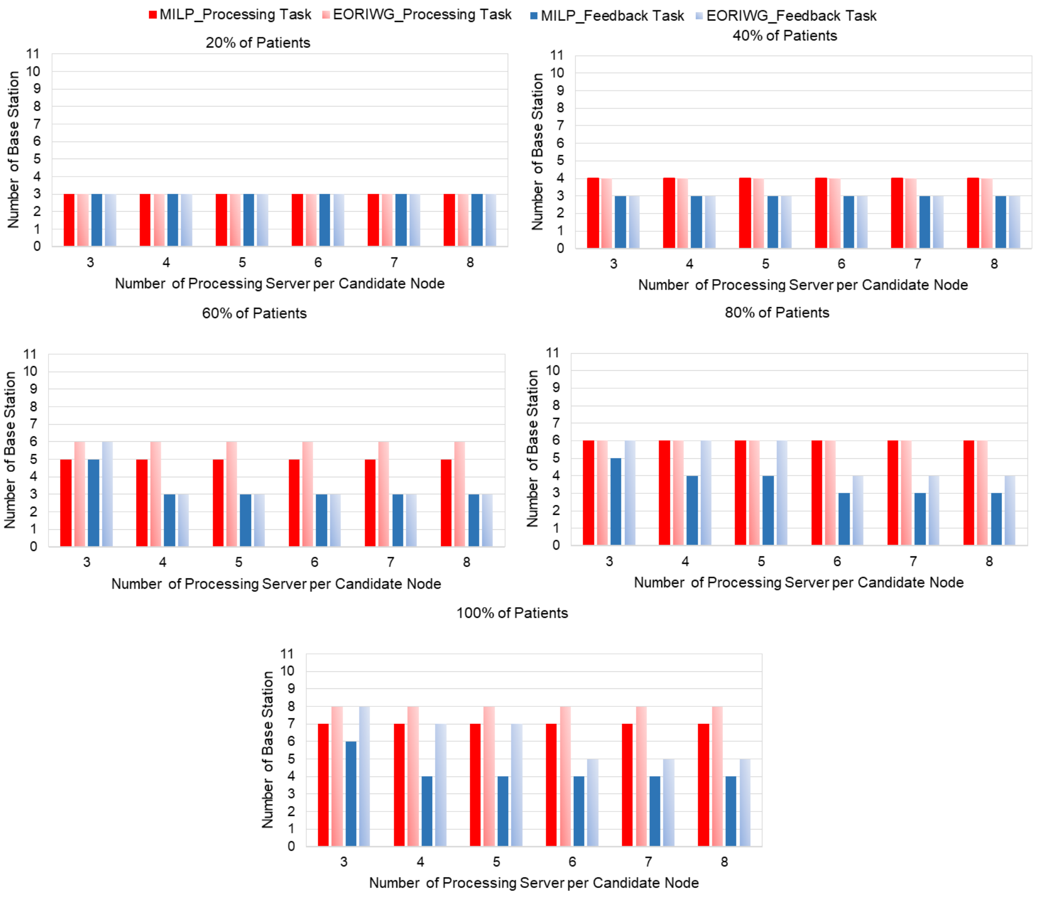


(b)

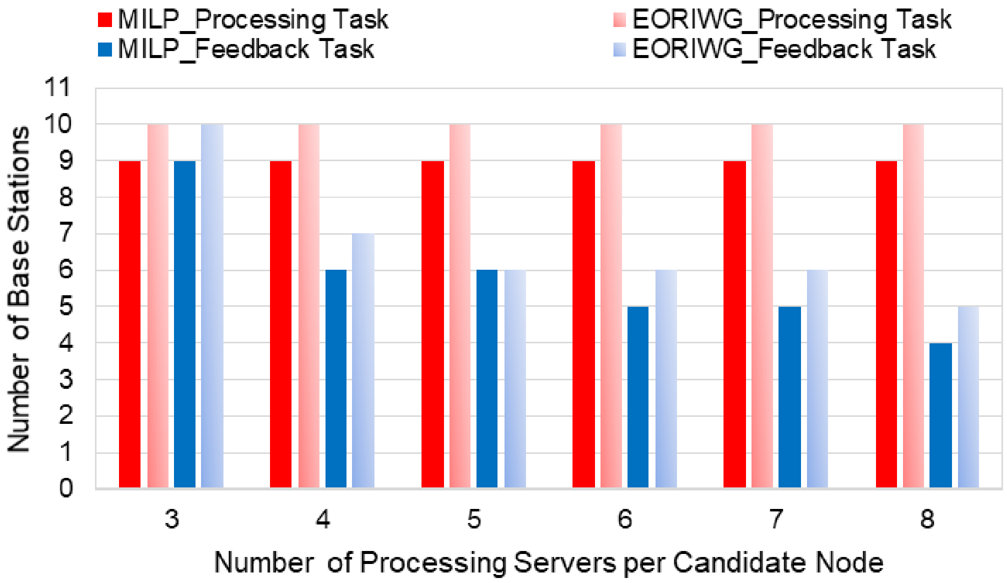
Figure 6.36: Total energy consumption of both networking equipment and processing for the MILP model and the EORIWG heuristic for (a) the ECG monitoring application with different percentages of the total number of patients (b) fall monitoring application at 100% of the total number of patients, for different number of processing servers per candidate node

Table 6.7: Average optimisation gaps between the MILP model and EORIWG heuristic for ECG and fall monitoring applications for different percentages of patients

Percentage of Patients	Type of Monitoring	Percentage of Patients				
		20%	40%	60%	80%	100%
Total Energy	ECG	0%	0%	0.17%	0.42%	0.44%
	Fall	-	-	-	-	2.56%
Network Energy	ECG	0%	0%	12.25%	32.36%	35.31%
	Fall	-	-	-	-	26.58%
Processing Energy	ECG	0%	0%	0%	0%	0%
	Fall	-	-	-	-	0%



(a)



(b)

Figure 6.37: Number of base stations used to serve the processing and feedback tasks for the MILP model and the EORIWG heuristic for (a) the ECG monitoring application with different percentages of the total number of patients, (b) the fall monitoring application at 100% of the total number of patients, for different number of processing servers per candidate node

6.10.2 EORIG heuristic results

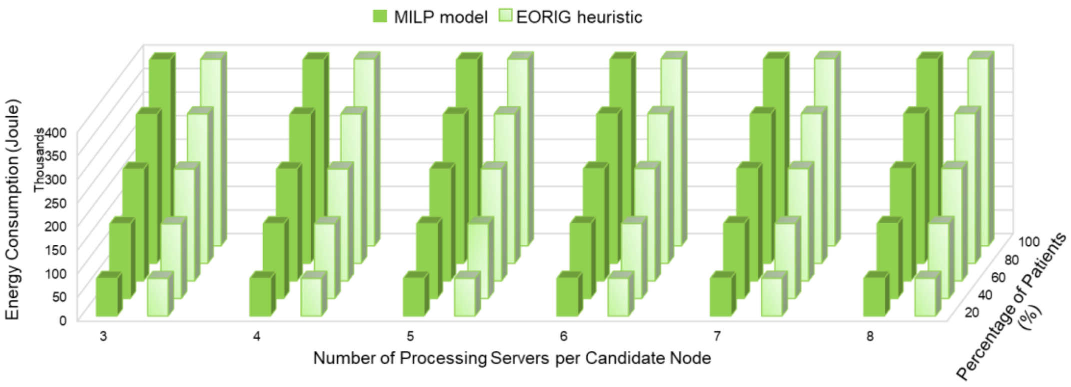
The results in Figure 6.38-(a) for ECG monitoring applications show that the total energy consumption of EORIG heuristic is equal to that produced by the MILP model at demand levels of 20% and 40% for all the different numbers of processing servers per candidate node. This is mainly due to the ability to utilise the minimum number of primary and secondary processing servers and number of candidate nodes to place the processing servers that are built in the EORIG heuristic while assigning the patients to the processing servers.

Also, as the size of demand is small, the same number of networking equipment is utilised in both EORIG heuristic and MILP model.

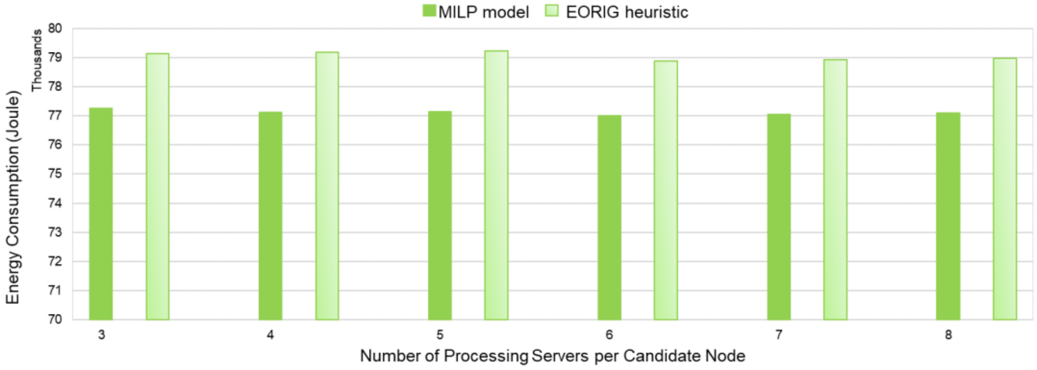
Figure 6.38-(a) also shows that the total energy consumption of the EORIG heuristic is higher than that produced by the MILP optimisation model with an average increase of 0.17%, 0.39% and 0.39% when the demand levels are 60%, 80% and 100%, respectively. The high energy consumed in EORIG heuristic at demand levels of 60% and 100% is due to the high number of utilised base stations to send the processing and feedback traffic as shown in Figure 6.39-(a). Note that, the base stations in cluster 1 and cluster 2 are used to serve the processing traffic due to the limitation of the connection between the clinics and the base stations. Also, at 80% and 100% of the maximum demand level, the higher energy consumed in the EORIG heuristic is due to the utilisation of the centre aggregation switch (CAS) to relay the processing traffic between the clusters to the processing servers. This is due to the same reason as explained in Section 6.9.1.

Meanwhile, Figure 6.38-(b) for fall monitoring applications, shows that the total energy consumed in EORIG heuristic is higher than that reported by the MILP optimisation model with an average increase of 2.51%. The higher energy consumed in the EORIG heuristic is due to the higher number of utilised networking equipment, including the base stations as shown in Figure 6.39-(b) in the EORIG heuristic compared to the MILP model as explained above. Also, this is due to the utilisation of the centre aggregation switch in the EORIG heuristic to relay the raw health data traffic between the clusters.

We also evaluate the computational time to run the EORIG heuristic and the MILP model for 100% of patients and three processing servers per candidate node. The results show that the EORIGW heuristic running on a normal PC with 3.2 GHz CPU and 16 GB RAM took 47 sec and 101 sec to finish for ECG and fall monitoring applications, respectively. Meanwhile, the MILP model running on a high-performance computing (HPC) cluster with a 12 core CPU and 64 GB RAM for ECG and fall monitoring applications was manually stopped after 47 hours and 24 hours, respectively.



(a)

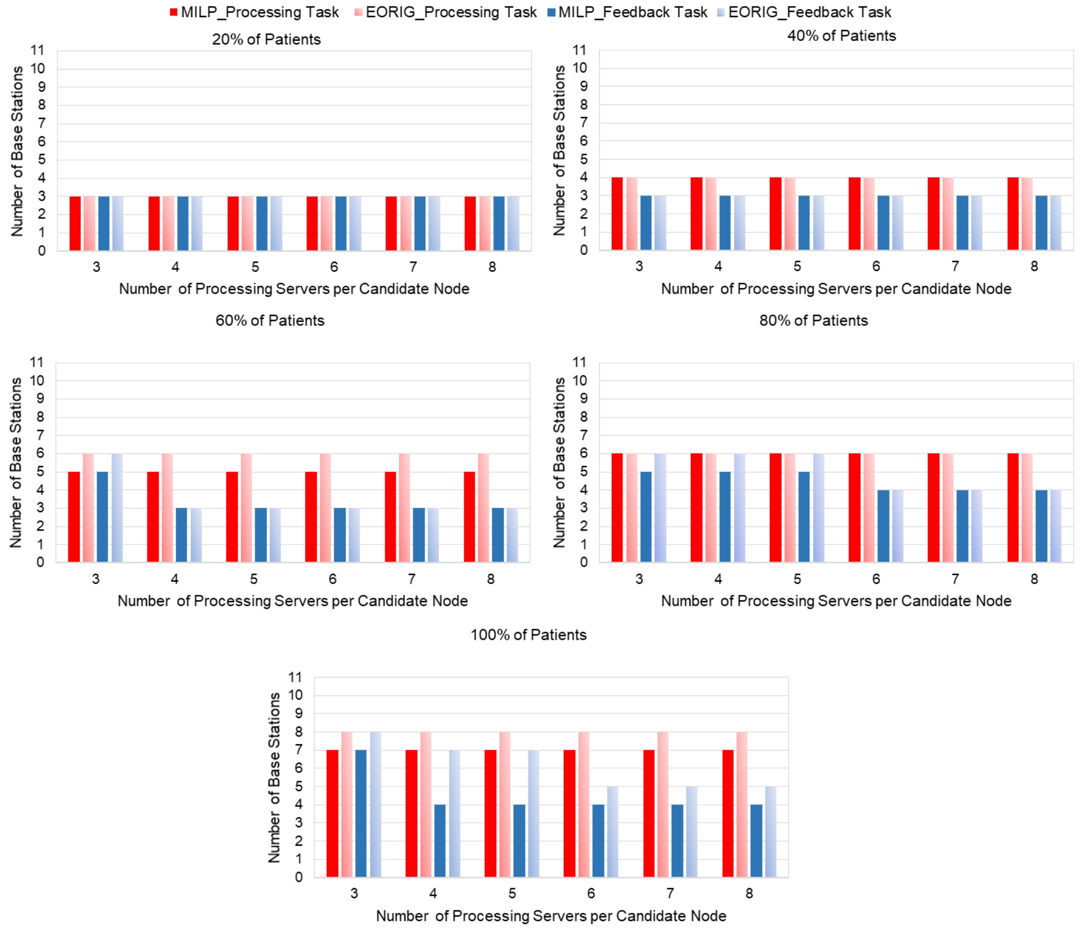


(b)

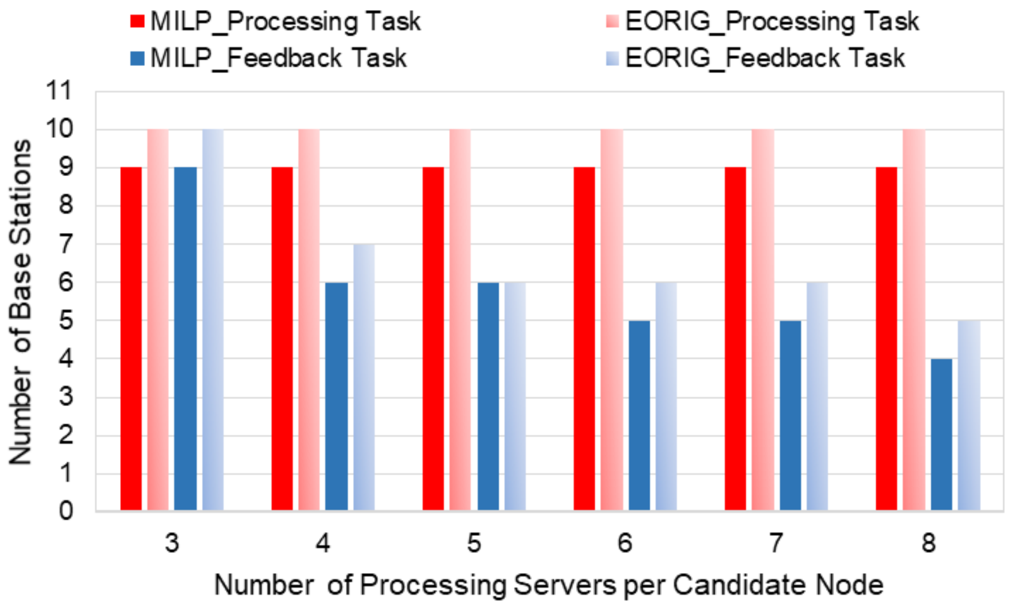
Figure 6.38: Total energy consumption of both networking equipment and processing for the MILP model and the EORIG heuristic for (a) ECG monitoring applications with different percentage of patients (b) fall monitoring applications at 100% of patients, for different number of processing servers per candidate node

Table 6.8: Average optimisation gaps between the MILP model and EORIG heuristic for ECG and fall monitoring applications

Percentage of Patients	Type of Monitoring	Percentage of Patients				
		20%	40%	60%	80%	100%
Total Energy	ECG	0%	0%	0.17%	0.39%	0.39%
	FALL	-	-	-	-	2.51%
Network Energy	ECG	0%	0%	11.95%	29.01%	30.11%
	FALL	-	-	-	-	25.8%
Processing Energy	ECG	0%	0%	0%	0%	0%
	FALL	-	-	-	-	0%



(a)



(b)

Figure 6.39: Number of base stations used to serve the processing and feedback tasks for the MILP model and the EORIG heuristic for (a) ECG monitoring application with different percentage of patients, (b) fall monitoring application at 100% of patients, for different number of processing servers per candidate node

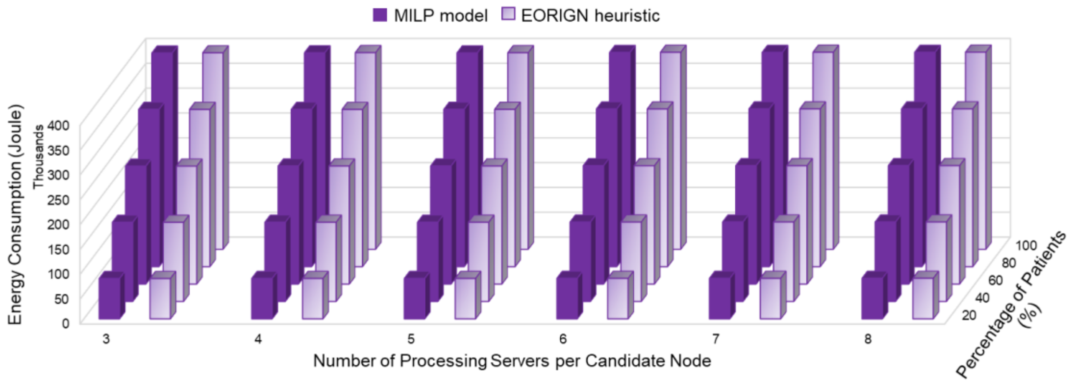
6.10.3 EORIGN heuristic results

The results in Figure 6.40-(a) for ECG monitoring applications show that the total energy consumption of EORGN heuristic is equal to the energy consumption reported by the MILP optimisation model at demand levels of 20%, 40%, 60%, and 80% for all number of processing servers per candidate node. This is due to the same amount of utilised networking equipment and processing servers in both models. Figure 6.40-(a) also shows that, at a demand level of 100%, the total energy consumption of the EORIGN heuristic is slightly higher than the MILP model with an average difference of about 0.1%. This is due to the limited number of connections between the base stations and the clinics in each cluster. Hence resulting in the utilisation of a higher number of base stations in the EORIGN heuristic, as shown in Figure 6.41-(a) to serve the processing traffic without maximising the utilisation of its resources.

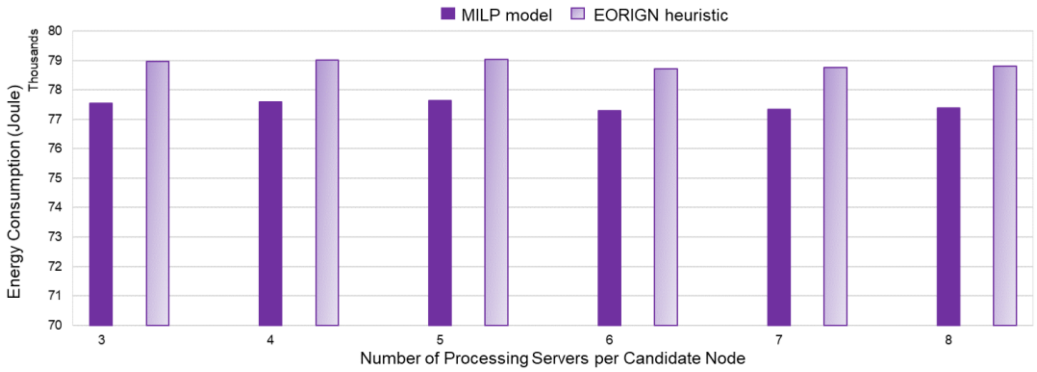
Meanwhile, for fall monitoring applications, Figure 6.40-(b) shows that the total energy consumption in the EORIGN heuristic is higher than the MILP model with an average energy increase of 1.83% in the EORIGN heuristic compared to the MILP. The increase in the energy consumption reported by the EORIGN heuristic is due to the higher number of base stations used to serve both processing and feedback traffic as shown in Figure 6.41-(b).

Table 6.9 also shows that, for all the considered demands and number of processing servers per candidate node, there are no optimisation gaps between the processing energy of the EORIGN heuristic and MILP model for the ECG and fall monitoring applications. This is due to the minimal number of processing servers utilised in the EORIGL heuristic, hence the same processing energy is consumed in both EORIGL heuristic and MILP model.

We also evaluate the computational time needed to run the EORIGN heuristic and the MILP model for 100% of patients and three processing servers per candidate node. The results show that, the EORIGN heuristic running on a normal PC with 3.2 GHz CPU and 16 GB RAM took 18 sec and 15 sec to finish for ECG and fall monitoring applications, respectively. Meanwhile the MILP model running on a high performance computing (HPC) cluster with a 12 core CPU and 64 GB RAM for ECG monitoring applications is manually stopped after 24 hours while for fall monitoring applications it took 192 sec to finish.



(a)

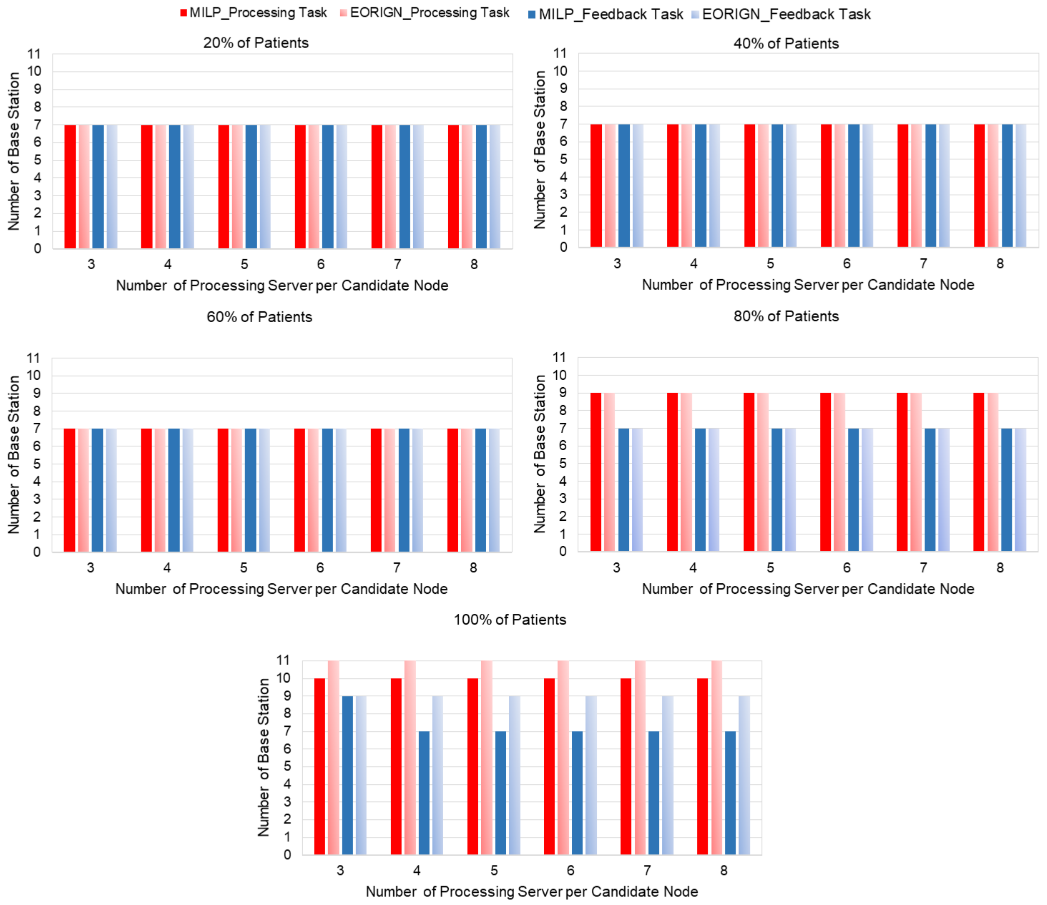


(b)

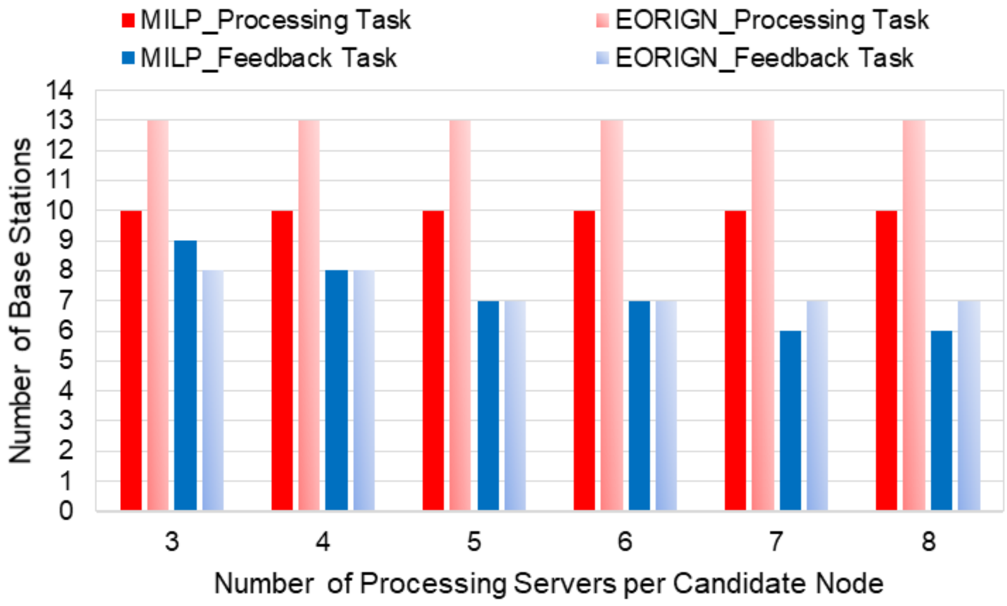
Figure 6.40: Total energy consumption of networking equipment and processing for the MILP model and the EORIGN heuristic (a) ECG monitoring application with different percentages of the total number of patients (b) fall monitoring application at 100% of the total number of patients, for different number of processing servers per candidate node

Table 6.9: Optimisation gaps between the MILP model and EORIGN heuristic for ECG and fall monitoring applications

Percentage of Patients	Type of Monitoring	Percentage of Patients				
		20%	40%	60%	80%	100%
Total energy	ECG	0%	0%	0%	0%	0.10%
	FALL	-	-	-	-	1.83%
Network energy	ECG	0%	0%	0%	0%	7.79%
	FALL	-	-	-	-	18.08%
Processing energy	ECG	0%	0%	0%	0%	0%
	FALL	0%	0%	0%	0%	0%



(a)



(b)

Figure 6.41: Number of base stations used to serve the processing and feedback tasks for the MILP model and the EORIGN heuristic for (a) ECG monitoring application with different percentages of the total number of patients, (b) fall monitoring application at 100% of the total number of patients, for different number of processing servers per candidate node

6.11 Conclusions

In this chapter, a resilient energy efficient fog computing infrastructure for health monitoring applications was proposed. The infrastructure is designed to be resilient against server failures under two scenarios; without geographical constraints and with geographical constraints and resilient against both server and network failures with geographical constraints and link and node disjoint resilience, respectively. Two types of health applications are considered to evaluate the performance of the proposed resilient infrastructure separately, which are ECG and fall monitoring applications. The patients from the clinics will send the required health data, for processing, analysis and decision making at both primary and secondary processing servers. A Mixed Integer Linear Programming (MILP) model, is used to optimise the number and locations of the primary and secondary processing servers, so that the energy consumption of both networking equipment and processing are minimised.

The results show that, considering a scenario for server protection without geographical constraints compared to the non-resilient scenario, incurs a high energy penalty of networking equipment for high demand. However, this penalty can be reduced by increasing the number of processing servers allowed at each candidate node, as more processing servers can be placed at each candidate node, hence reducing the amount of networking equipment (i.e. Ethernet switches) utilised. The results also reveal that the energy consumption of processing in the resilient scenario without any geographical constraints, is higher than that of the non-resilient scenario, as the latter scenario does consider secondary processing servers.

Meanwhile, for ECG monitoring applications, increasing the level of resilience to consider geographical constraints at a low level of demand (i.e. 20%), has resulted in the highest energy penalty, compared to the higher levels of demand. This is because more nodes are utilised, to place the processing servers under the geographic constraints. However, when the demand level increases from 40% to 100%, increasing the level of resilience does not incur an energy penalty, and this depends on the number of processing servers allowed at each candidate node. Also, increasing the number of processing servers per candidate node at a demand level between 40% and 100%, can either decrease or increase the energy penalty. The increase in the energy penalty, is because of the reduction in the number of candidate nodes needed to place the processing servers in the resilient scenario, without geographical constraints. On the other hand, the decrease in the energy penalty, is because of the reduced number of candidate nodes needed to place the processing servers in the resilient scenario, with geographical constraints. However, the energy penalty due to considering geographical constraints at a demand level equal to or more than 40%, is less than 7%. The results also show that the same energy of processing is consumed in both resilient scenarios, for all processing servers per candidate node. This is because the same number of servers are used in both scenarios, as the patients were optimally consolidated in the processing servers. Also, the same patterns of energy penalty occur in fall monitoring applications, at a demand level of 100%.

The results also show that increasing the level of resilience to consider the geographical constraint for server protection and link and node disjoint resilience for network protection compared to only geographic constraints gives the same energy consumption of processing, while increasing the energy consumption of networking equipment. The increase in energy of networking is the penalty for having a higher level of resilience. The results indicate that considering disjoint link and node resilience has resulted in a low network energy penalty at high demands for both ECG and fall monitoring applications due to the activation of a large part of the network in any case due to the demands. Also, for both ECG and fall monitoring applications, increasing the number of processing servers at each candidate node can reduce the energy penalty of the network at high demand levels.

Three heuristic models, EORIWG heuristic, EORIG heuristic, and EORIGN heuristic were also developed for real-time solution of the resilient scenario without geographical constraints, with geographical constraints, and both with geographical constraints and link and node disjoint, respectively. The results show that the performance of both EORIWG and EORIG heuristic models are the same as the MILP models at low demand levels (i.e. 20%) while for the EORIGN heuristic, the performance are the same as MILP model at demand level of 20% to 80%, under ECG monitoring applications. However, increasing the level of demand has resulted in an increase in the energy consumption of the networking equipment in EORIWG, EORIG and EORIGN heuristics compared to the MILP models. This is mainly due to the high number of networking equipment utilised in the networks. However, the total optimisation gaps between the heuristics and MILP models for both monitoring applications is less than 3%.

Chapter 7

Summary of contributions and future work

This chapter provides a summary of the main contributions that have been presented in this thesis. It also suggests possible future research directions in the area of energy efficiency of health monitoring applications leveraging the use of fog computing.

7.1 Summary of contributions

The limited computation and storage capabilities of a wireless body sensor have led to the need for cloud computing for health monitoring applications. However, the massive data transfer to the cloud contributes to high latency. Therefore, fog computing is a potential solution that can overcome the limitations at the cloud network. The increasing ageing population and chronic diseases have increased the demand for remote health monitoring services. With this rising demand for fog-based remote healthcare services, the energy consumption in the transport network has become a significant issue.

To address the energy issues, the first contribution in this thesis was to propose a network framework for energy-efficient health monitoring applications leveraging fog computing realised in a network architecture based on Gigabit Passive Optical Networks (GPON).

Energy-efficient fog computing (EEFC) was proposed in Chapter 3, and a MILP optimisation model was developed to minimise the energy consumption of networking equipment and processing. The number and locations of the processing servers (i.e. fog servers) were optimised in the access network. In Chapter 4, the performance of the EEFC model (fog approach) was evaluated using low data rate ECG monitoring applications. A maximum network saving of 83.1% was achieved by the EEFC model when compared to the conventional approach (i.e. energy-efficient cloud computing, EECC model), where the processing is performed at the cloud. Additionally, the total saving of both networking and processing in the EEFC model was 35.7% compared to the EECC model. For real-time implementation, the Energy Optimised Fog Computing (EOFC) heuristic algorithm was developed. The results show the EOFC heuristic works as good as the MILP model. This is due to the ability to use the minimum number of processing servers, the minimum number of candidate nodes to place the processing servers and the minimum number of networking equipment (i.e. base stations) that are built into the EOFC heuristic while assigning the patients from clinics to the processing servers. Also, the computational time for EOFC heuristic is lower than the MILP model. The investigation was taken further to show the energy efficiency of fog computing for health monitoring applications, considering a reduction in the idle power of the equipment, increasing traffic and different access network (i.e. Ethernet). The results show that the EEFC model and EOFC heuristic always has the lowest energy consumption compared to the conventional approach. However, allowing only one processing server at each candidate node at the access layer has resulted in high energy consumption of both networking equipment and processing compared to the EECC model. The high energy consumed in the EEFC model is due to the limited link capacities at the access network to consolidate patients in the processing servers; hence, it increases the total utilised processing servers and networking equipment in the fog approach

The results also reveal that deploying fog computing under a GPON network is more energy-efficient than in the Ethernet network. It is worth noting that the OLT (i.e. aggregation switch for Ethernet access network) is always chosen to place the processing servers. However, due to the increasing number of patients and the limited number of processing servers that can be served in a single node, this has resulted in placing the remaining processing servers at the ONU.

The energy efficiency of fog computing to serve high data rate for fall monitoring applications was also investigated in Chapter 5. Two scenarios were considered. The first involved limiting the number of patients that can be served in a single processing server. The second limited the number of processing servers that are allowed at each candidate node, while a single processing server can serve 20% of the maximum patients considered in the network. The same MILP model in Chapter 3 was used with additional constraints to meet the requirements of the second scenario due to limiting the number of processing servers allowed at each candidate node. The results reveal that a total energy saving of 37.7% is achieved in the EEFC model compared to the EECC model, when 20% of the maximum patients can be served in a single processing server. Furthermore, the total energy consumption of networking equipment and processing in both fog (EEFC model) and cloud (EECC model) reduced significantly when more patients can be served in a single processing server as less processing servers are utilised. There was a 52.2% total energy saving achieved in the EEFC model compared to the EECC model when all patients can be served by a single processing server. Meanwhile, limiting the number of processing servers that can be accommodated at each candidate node resulted in high energy consumption of networking equipment and processing. This is because more networking devices (i.e. Ethernet switches) are utilised to place the processing servers. Additionally, the results show that the OLT is always used to place the processing servers, as it is the nearest central location to the patients. However, due to the limited number of servers that can be accommodated in a single node, the ONU is used to place the remaining processing servers that cannot be allocated at the OLT. The performance of EOFC heuristic was also investigated, and the results show that it works as good as the MILP model in both scenarios with less than 3% optimisation gap. The low optimisation gap is due to the same reasons as explained above for ECG monitoring applications. In addition, the computational time for EOFC heuristic is lower than the MILP models.

A resilient energy-efficient fog computing infrastructure for health monitoring applications for server protections was investigated in Chapter 6. The MILP model was developed to optimise the number and location of primary and secondary processing servers to minimise the total energy consumption of networking equipment and processing under ECG and fall monitoring applications, separately. The MILP model is tested under two-levels of resilience without geographical constraints and with geographical constraints. The results reveal that considering resilience has resulted in 15% - 86% network energy penalty under low data rate for ECG monitoring applications, depending on the level of demand, the energy penalty with high data rate for fall monitoring applications is approximately 93% at high demand.

Moreover, for both applications, the added resilience increased the energy consumption associated with processing by 50%. The increase in the energy of network and processing is because the resilient scenario without geographical constraints has increased the total traffic and the number of processing servers. However, the energy penalty can be reduced when more processing servers are accommodated at each candidate node, as fewer Ethernet switches are used to place the processing servers. Adding geographical constraints for server protection (i.e. primary and secondary processing servers have to be node disjoint) does not add any significant energy consumption of networking equipment (i.e. less than 10%), while the same energy of processing is consumed for both ECG and fall monitoring applications. Furthermore, increasing the number of processing servers can reduce the energy penalty in the more resilient scenario.

Since geographical constraints do not significantly increase the energy consumption, therefore, Chapter 6 also considered the protection of the servers with geographical constraints with additional network protection with link and node disjoint resilience. The same MILP model for resilient scenario with geographical constraints is used with additional constraints for network protection. The results show that considering server protection with geographical constraints and network protection with link and node disjoint resilience resulted in 17% - 86% network energy penalty under low data rate for ECG monitoring applications. This penalty depends on the demand level. Meanwhile, the energy penalty is approximately 6.3% when a high data rate for fall application is considered at high demand. However, the energy penalty of the network can be reduced by increasing the number of processing servers that can be served at each candidate node. The reduction in energy penalty occurs because increasing the number processing servers at each candidate node can reduce the amount of Ethernet switches to place the processing servers. Also, the same energy of processing is consumed regardless of the increase in level of resilience. The results also show that, the performance of the developed heuristic for all resilient scenarios approaches that of the MILP results with less than 3% optimisation gap. This is due to the same reasons as explained for the EOFC heuristic. Also, the computational time to run the heuristic models are lower than the MILP models.

As conclusions, this work has shown that optimising the number and location of processing servers at fog (i.e. network edge) can reduce the energy consumption of networking equipment and processing compared to the conventional approach where the processing servers are located at the cloud. The low energy consumption of the networking equipment and processing in the proposed approach can reduce the cost of operations for services providers. The proposed fog computing approach can also be used for other applications by changing the parameters related to that application. However, this only limited to the applications that have specific time constraints to determine the parameter inputs. Also note that different parameters (i.e. time constraints, processing and analysis time of the health data) used in the model will impact the performance of both fog computing approach and cloud computing approach in terms of energy consumption of the networking equipment and processing. In addition, this work also studies the impact of increasing the level of resilience in the fog computing infrastructure and proposed a solution to reduce the energy consumption of networking equipment and processing by increasing the number of processing servers per candidate node. We also developed a heuristic model for each MILP model to validate the MILP operation and to deliver a real-time solution based on the insights from the MILP result. The results show that the performance of the heuristics is equal to or approaching the MILP results with less than 3% optimisation gaps. In addition, the computational time for all developed heuristics is lower than the MILP models. Therefore, the developed heuristics can be used for other applications with specific time constraints as explained above and are not limited to health monitoring applications.

7.2 Future directions

This thesis has investigated the use of fog computing for Internet of Things applications (i.e. health monitoring applications) to improve the energy efficiency of networking equipment and processing. Various MILP models and heuristic-based algorithms have been proposed and shown to improve the energy performance of networking equipment and processing significantly. Although the energy of networking equipment and processing increase with the increasing level of resilience, however, these energies are reduced by increasing the number of servers that can be served at each candidate node. These investigations have helped identify the following future research directions that could be explored:

1. **Energy-efficient network slicing for healthcare applications**

In this thesis, we investigated two types of health monitoring applications; ECG monitoring applications and fall monitoring applications running separately. Therefore, it is valuable to study different healthcare applications that require different processing time and have different delay tolerances when running at the same time. A model could be developed in this scenario to carry out network embedding while minimising the energy consumption of networking equipment and processing.

2. Resilient virtualised infrastructure for healthcare

In this thesis, the impact of increasing the level of resilience for server protection and network protection on the energy consumption of both networking equipment and processing has been investigated. Therefore, in addition to the idea of network embedding for multiple healthcare applications, it is valuable to explore the potential to reduce the energy consumption of the networking equipment and processing by allowing one network slice to share the processing and networking infrastructure of the other network slice for protection. A model could be developed to carry out network embedding considering the proposed resilient scenarios in this thesis, while minimising the energy consumption of both networking equipment and processing.

References

[1] X. Dong, T. El-Gorashi, and J. Elmirghani, “Energy Efficient Optical Networks with Minimized Non-Renewable Power Consumption,” *J. Networks*, vol. 7, no. 5, pp. 821–831, 2012.

[2] C. Bianco, F. Cucchietti, and G. Griffa, “Energy consumption trends in the Next Generation Access Network - A Telco perspective,” *INTELEC, Int. Telecommun. Energy Conf.*, pp. 736–742, 2007.

[3] X. Dong, T. El-Gorashi, and J. M. H. Elmirghani, “Green IP Over WDM Networks With Data Centres,” *2011 13th Int. Conf. Transparent Opt. Networks*, vol. 29, no. 12, pp. 1–8, 2011.

[4] R. Bolla, R. Bruschi, F. Davoli, and F. Cucchietti, “Energy Efficiency in the Future Internet: A Survey of Existing Approaches and Trends in Energy-Aware Fixed Network Infrastructures,” *Commun. Surv. Tutorials, IEEE*, vol. 13, no. 2, pp. 223–244, 2011.

[5] E. Farnworth and J. C. Castilla-rubio, “SMART 2020 : Enabling the low carbon economy in the information age,” 2020.

[6] A. Abayomi-Alli, A. J. Ikuomola, O. A. Aliyu, and O. Abayomi-Alli, “Development of a Mobile Remote Health Monitoring system – MRHMS,” *African J. Comput. ICT African J. Comput. ICT Ref. Format Afr J. Comp ICTs*, vol. 7, no. 4, pp. 14–22, 2014.

[7] U. Nations, “World Population Ageing 2013,” *Dep. Econ. Soc. Aff. Popul. Div.*, pp. 1–114, 2013.

[8] S. Stewart, N. Murphy, A. Walker, A. Mcguire, and J. J. V Mcmurray, “Cost of an emerging epidemic: an economic analysis of atrial fibrillation

in the UK," *Hear. BMJ Journals*, vol. 90, pp. 286–293, 2004.

[9] Z. Pang, "Technologies and Architectures of the Internet-of-Things (IoT) for Health and Well-being," Royal Institute of Technology, 2013.

[10] P. L. and T. S. I. Azimi, A. Anzanpour, A. M. Rahmani, "Medical Warning System Based on Internet of Things Using Fog Computing," in *2016 International Workshop on Big Data and Information Security (IWBIS)*, 2016, pp. 19–24.

[11] N. K. and S. Z. K. Kaur, T. Dhand, "Container as a Service at the Edge : Trade - off between Energy Efficiency and Service Availability at Fog Nano Data Centers," *IEEE Wirel. Commun.*, vol. 24, no. June, pp. 48–56, 2017.

[12] T. N. Gia, M. Jiang, A. M. Rahmani, T. Westerlund, P. Liljeberg, and H. Tenhunen, "Fog computing in healthcare Internet of Things: A case study on ECG feature extraction," in *IEEE International Conference on Computer and Information Technology; Ubiquitous Computing and Communications; Dependable, Autonomic and Secure Computing; Pervasive Intelligence and Computing*, 2015, pp. 356–363.

[13] F. Jalali, K. Hinton, R. S. Ayre, T. Alpcan, and R. S. Tucker, "Fog Computing May Help to Save Energy in Cloud Computing," *IEEE J. Sel. Areas Commun.*, vol. 34, no. 5, pp. 1728–1739, 2016.

[14] L. F. Bittencourt, M. M. Lopes, I. Petri, and O. F. Rana, "Towards Virtual Machine Migration in Fog Computing," pp. 1–8, 2015.

[15] S. K. Datta, C. Bonnet, and J. Haerri, "Fog Computing architecture to enable consumer centric Internet of Things services," in *IEEE*

International Symposium on Consumer Electronics (ISCE), 2015, pp. 1–2.

[16] N. Constant, D. Borthakur, M. Abtahi, H. Dubey, and K. Mankodiya, “Fog-Assisted wIoT: A Smart Fog Gateway for End-to-End Analytics in Wearable Internet of Things,” in *arXiv preprint arXiv:170108680*, 2017, pp. 1–5.

[17] Q. Zhang, X. Zhang, Q. Zhang, W. Shi, and H. Zhong, “Firework: Big data sharing and processing in collaborative edge environment,” *Proc. - 4th IEEE Work. Hot Top. Web Syst. Technol. HotWeb 2016*, pp. 20–25, 2016.

[18] T. H. Luan, L. Gao, Z. Li, Y. Xiang, and L. Sun, “Fog Computing: Focusing on Mobile Users at the Edge,” *eprint arXiv:1502.01815*, 2015.

[19] S. Yi, C. Li, and Q. Li, “A Survey of Fog Computing: Concepts, Applications and Issues,” in *Proceedings of the 2015 Workshop on Mobile Big Data - Mobidata ’15*, 2015, pp. 37–42.

[20] M. Aazam and E. N. Huh, “Fog computing micro datacenter based dynamic resource estimation and pricing model for IoT,” *Proc. - Int. Conf. Adv. Inf. Netw. Appl. AINA*, vol. 2015-April, pp. 687–694, 2015.

[21] M. Aazam and E. N. Huh, “E-HAMC: Leveraging Fog computing for emergency alert service,” in *IEEE International Conference on Pervasive Computing and Communication Workshops, PerCom Workshops 2015*, 2015, pp. 518–523.

[22] O. Consortium and A. Working, “OpenFog Reference Architecture for Fog Computing,” no. February, pp. 1–162, 2017.

[23] A. Modarresi, S. Gangadhar, and J. P. G. Sterbenz, “A framework for improving network resilience using SDN and fog nodes,” *Proc. 2017 9th Int. Work. Resilient Networks Des. Model. RNDM 2017*, pp. 1–7, 2017.

[24] J. M. H. Elmirghani *et al.*, “GreenTouch GreenMeter Core Network Energy-Efficiency Improvement Measures and Optimization,” *J. Opt. Commun. Netw.*, vol. 10, no. 2, p. A250, 2018.

[25] L. Nonde, T. E. H. El-gorashi, and J. M. H. Elmirghani, “Energy Efficient Virtual Network Embedding for Cloud Networks,” *J. Light. Technol.*, vol. 33, no. 9, pp. 1828–1849, 2015.

[26] A. Q. Lawey, T. E. H. El-Gorashi, and J. M. H. Elmirghani, “BitTorrent Content Distribution in Optical Networks,” *J. Light. Technol.*, vol. 32, no. 21, pp. 4209–4225, 2014.

[27] N. I. Osman, T. El-Gorashi, L. Krug, and J. M. H. Elmirghani, “Energy-efficient future high-definition TV,” *J. Light. Technol.*, vol. 32, no. 13, pp. 2364–2381, 2014.

[28] X. Dong, T. El-gorashi, and J. M. H. Elmirghani, “On the energy efficiency of physical topology design for IP over WDM networks,” *J. Light. Technol.*, vol. 30, no. 12, pp. 1931–1942, 2012.

[29] X. Dong, T. El-Gorashi, and J. M. H. Elmirghani, “Green IP over WDM networks with data centers,” *J. Light. Technol.*, vol. 29, no. 12, pp. 1861–1880, 2011.

[30] X. Dong, T. El-Gorashi, and J. M. H. Elmirghani, “IP Over WDM Networks Employing Renewable Energy Sources,” *J. Light. Technol.*, vol. 29, no. 1, pp. 3–14, 2011.

[31] M. Musa, T. Elgorashi, and J. Elmirghani, “Energy efficient survivable IP-Over-WDM networks with network coding,” *J. Opt. Commun. Netw.*, vol. 9, no. 3, pp. 207–217, 2017.

[32] A. M. Al-Salim, A. Q. Lawey, T. E. H. El-Gorashi, and J. M. H. Elmirghani, “Energy Efficient Big Data Networks: Impact of Volume and Variety,” *IEEE Trans. Netw. Serv. Manag.*, vol. 15, no. 1, pp. 458–474, 2018.

[33] A. M. Rahmani *et al.*, “Exploiting smart e-Health gateways at the edge of healthcare Internet-of-Things: A fog computing approach,” *Futur. Gener. Comput. Syst.*, vol. 78, pp. 641–658, 2018.

[34] S. Tayeb, S. Latifi, and Y. Kim, “A survey on IoT communication and computation frameworks: An industrial perspective,” *2017 IEEE 7th Annu. Comput. Commun. Work. Conf. CCWC 2017*, vol. 1301726, pp. 1–6, 2017.

[35] J. Gubbi, R. Buyya, S. Marusic, and M. Palaniswami, “Internet of Things (IoT): A vision, architectural elements, and future directions,” *Futur. Gener. Comput. Syst.*, vol. 29, pp. 1645–1660, 2013.

[36] B. Farahani, F. Firouzi, V. Chang, M. Badaroglu, N. Constant, and K. Mankodiya, “Towards fog-driven IoT eHealth: Promises and challenges of IoT in medicine and healthcare,” *Futur. Gener. Comput. Syst.*, vol. 78, pp. 659–676, 2018.

[37] S. F. Abedin, M. G. R. Alam, R. Haw, and C. S. Hong, “A system model for energy efficient green-IoT network,” in *2015 International Conference on Information Networking (ICOIN)*, 2015, pp. 177–182.

[38] S. (2012). Khan, R., Khan, S. U., Zaheer, R., & Khan, "Applications and Key Challenges Future Internet : The Internet of Things Architecture , Possible Applications and Key Challenges," in *10th International Conference on Frontiers of Information Technology (FIT)*, 2012, pp. 257–260.

[39] GSMA, "Understanding the Internet of Things (IoT)," 2014.

[40] G. C. Jana and S. Banerjee, "Enhancement of QoS for fog computing model aspect of robust resource management," *2017 Int. Conf. Intell. Comput. Instrum. Control Technol.*, pp. 1462–1466, 2017.

[41] F. Fernandez and G. C. Pallis, "Opportunities and challenges of the Internet of Things for healthcare: Systems engineering perspective," *Proc. 2014 4th Int. Conf. Wirel. Mob. Commun. Healthc. - "Transforming Healthc. Through Innov. Mob. Wirel. Technol. MOBIHEALTH 2014*, pp. 263–266, 2015.

[42] H. of C. L. Ana Tavares Lattibeaudiere, "Connected Living Latin America Summit," *GSMA Connected Living*, pp. 1–19, 2012.

[43] M. Aazam, P. Hung, and E. Huh, "Smart gateway based communication for cloud of things," in *Intelligent Sensors, Sensor Networks and Information Processing*, 2014, pp. 1–6.

[44] M. Aazam and E. N. Huh, "Fog computing and smart gateway based communication for cloud of things," *Proc. - 2014 Int. Conf. Futur. Internet Things Cloud, FiCloud 2014*, pp. 464–470, 2014.

[45] N. Sultan, "International Journal of Information Management Making use of cloud computing for healthcare provision : Opportunities and

challenges," *Int. J. Inf. Manage.*, vol. 34, no. 2, pp. 177–184, 2014.

[46] A. Modarresi and J. P. G. Sterbenz, "Toward Resilient Networks with Fog Computing," in *9th International Workshop on Resilient Networks Design and Modeling (RNDM), Alghero*, 2017, pp. 1–7.

[47] S. M. R. Islam, D. Kwak, and H. Kabir, "The Internet of Things for Health Care : A Comprehensive Survey," *Access, IEEE*, vol. 3, pp. 678–708, 2015.

[48] T. McCue, "$117 Billion Market For Internet of Things In Healthcare By 2020," *Forbes*, 2015. [Online]. Available: https://www.forbes.com/sites/tjmccue/2015/04/22/117-billion-market-for-internet-of-things-in-healthcare-by-2020/#3885985669d9.

[49] A. Botta, W. De Donato, V. Persico, and A. Pescapé, "Integration of Cloud computing and Internet of Things: A survey," *Future Generation Computer Systems*, vol. 56. pp. 684–700, 2016.

[50] D. G. Korzun, A. V Borodin, I. A. Timofeev, I. V Paramonov, and S. I. Balandin, "Digital Assistance Services for Emergency Situations in Personalized Mobile Healthcare : Smart Space based Approach," in *SIBIRCON/SibMedInfo Digital*, 2015, pp. 62–67.

[51] M. A. Al-taee, W. Al-nuaimy, A. Al-ataby, and Z. J. Muhsin, "Mobile Health Platform for Diabetes Management Based on the Internet-of-Things," in *IEEE Jordan Conference on Applied Electrical Engineering and Computing Technologies (AEECT) Mobile*, 2015, pp. 1–4.

[52] P. Ray, "Home Health Hub Internet of Things (H3 IoT): An architectural framework for monitoring health of elderly people," in *International*

Conference on Science, Engineering and Management Research (ICSEMR 2014), 2014, pp. 3–5.

[53] R. Mahmud, F. L. Koch, and R. Buyya, "Cloud-Fog Interoperability in IoT-enabled Healthcare Solutions," *Proc. 19th Int. Conf. Distrib. Comput. Netw. - ICDCN '18*, pp. 1–10, 2018.

[54] Y. YIN, Y. Zeng, X. Chen, and Y. Fan, "The internet of things in healthcare: An overview," *J. Ind. Inf. Integr.*, vol. 1, pp. 3–13, 2016.

[55] S. S. Bull, B. Gaglio, H. G. McKay, and R. E. Glasgow, "Harnessing the potential of the internet to promote chronic illness self-management: diabetes as an example of how well we are doing.," *Chronic Illn.*, vol. 1, no. 2, pp. 143–155, 2005.

[56] J. Granados, A. Rahmani, P. Nikander, P. Liljeberg, and H. Tenhunen, "Towards Energy-Efficient HealthCare : an Internet-of-Things Architecture Using Intelligent Gateways," in *Proceedings of the 4th International Conference on Wireless Mobile Communication and Healthcare - "Transforming healthcare through innovations in mobile and wireless technologies,"* 2014, pp. 279–282.

[57] P. Kakria, N. K. Tripathi, and P. Kitipawang, "A real-time health monitoring system for remote cardiac patients using smartphone and wearable sensors," *Int. J. Telemed. Appl.*, vol. 2015, pp. 1–11, 2015.

[58] Y. Shi, H. Wang, H. E. Roman, and S. Lu, "The Fog Computing Service for Healthcare," in *The Fog Computing Service for Healthcare*, 2015, pp. 1–5.

[59] T. Nguyen Gia *et al.*, "Low-cost fog-assisted health-care IoT system with

energy-efficient sensor nodes,” *2017 13th Int. Wirel. Commun. Mob. Comput. Conf. IWCMC 2017*, pp. 1765–1770, 2017.

[60] W. Tang, K. Zhang, J. Ren, Y. Zhang, and X. Shen, “Lightweight and Privacy-preserving Fog-assisted Information Sharing Scheme for Health Big Data,” in *GLOBECOM 2017 - 2017 IEEE Global Communications Conference*, 2017, pp. 1–6.

[61] N. A. Ali and A. Ain, “Internet of Nano-Things Healthcare Applications : Requirements , Opportunities , and Challenges,” in *International Workshop on Advances in Body-Centric Wireless Communications and Networks and Their Applications*, 2015, pp. 9–14.

[62] N. Jothi, N. Aini, A. Rashid, and W. Husain, “Data Mining in Healthcare – A Review,” *Procedia - Procedia Comput. Sci.*, vol. 72, pp. 306–313, 2015.

[63] O. Akrivopoulos, I. Chatzigiannakis, C. Tselios, and A. Antoniou, “On the Deployment of Healthcare Applications over Fog Computing Infrastructure,” *2017 IEEE 41st Annu. Comput. Softw. Appl. Conf.*, pp. 288–293, 2017.

[64] D. Guibert, J. Wu, S. He, M. Wang, and J. Li, “CC-fog: Toward content-centric fog networks for E-health,” *2017 IEEE 19th Int. Conf. e-Health Networking, Appl. Serv.*, pp. 1–5, 2017.

[65] L. Cerina, S. Notargiacomo, M. G. Paccanit, and M. D. Santambrogio, “A fog-computing architecture for preventive healthcare and assisted living in smart ambients,” *RTSI 2017 - IEEE 3rd Int. Forum Res. Technol. Soc. Ind. Conf. Proc.*, 2017.

[66] F. A. Kraemer, A. E. Braten, N. Tamkittikhun, and D. Palma, “Fog Computing in Healthcare – A Review and Discussion,” *IEEE Access*, vol. 3536, no. 2169, pp. 1–1, 2017.

[67] A.-E. T. Ali, Mai, Tuan Nguyen Gia, “Autonomous Patient / Home Health Monitoring powered by Energy Harvesting,” in *GLOBECOM 2017 - 2017 IEEE Global Communications Conference*, 2017, pp. 1–7.

[68] R. Craciunescu, A. Mihovska, and S. Halunga, “Implementation of Fog Computing for Reliable E- Health Applications,” in *49th Asilomar Conference on Signals, Systems and Computers*, 2015, pp. 5–9.

[69] O. Fratu, C. Pena, R. Craciunescu, and S. Halunga, “Fog computing system for monitoring Mild Dementia and COPD patients - Romanian case study,” in *12th International Conference on Telecommunications in Modern Satellite, Cable and Broadcasting Services, TELSIKS 2015*, 2015, pp. 123–128.

[70] P. Verma and S. K. Sood, “Fog Assisted-IoT Enabled Patient Health Monitoring in Smart Homes,” *IEEE Internet Things J.*, vol. 4662, no. c, pp. 1–8, 2018.

[71] Y. Cao, S. Chen, P. Hou, and D. Brown, “FAST: A fog computing assisted distributed analytics system to monitor fall for stroke mitigation,” in *IEEE International Conference on Networking, Architecture and Storage, NAS 2015*, 2015, pp. 2–11.

[72] A. T. Ozdemir, C. Tunc, and S. Hariri, “Autonomic fall detection system,” in *IEEE 2nd International Workshops on Foundations and Applications of Self* Systems, FAS*W 2017*, 2017, pp. 166–170.

[73] X. Masip-Bruin, E. Marin-Tordera, A. Gomez, V. Barbosa, and A. Alonso, “Will it be cloud or will it be fog? F2C, A novel flagship computing paradigm for highly demanding services,” in *Future Technologies Conference*, 2016, no. December, pp. 1129–1136.

[74] D. C. Yacchirema, D. Sarabia-Jacome, C. E. Palau, and M. Esteve, “A Smart System for sleep monitoring byintegrating IoT with big data analytics,” *IEEE Access*, vol. 6, pp. 35988–36001, 2018.

[75] I. Ishaq *et al.*, *IETF Standardization in the Field of the Internet of Things (IoT): A Survey*, vol. 2, no. 2. 2013.

[76] K. I. K. Wang, A. Rajamohan, S. Dubey, S. A. Catapang, and Z. Salcic, “A Wearable Internet of Things Mote with Bare Metal 6LoWPAN Protocol for Pervasive Healthcare,” *Proc. - 2014 IEEE Int. Conf. Ubiquitous Intell. Comput. 2014 IEEE Int. Conf. Auton. Trust. Comput. 2014 IEEE Int. Conf. Scalable Comput. Commun. Assoc. Sy*, pp. 750–756, 2015.

[77] A. Triantafyllou, P. Sarigiannidis, and T. D. Lagkas, “Network protocols, schemes, and mechanisms for internet of things (IoT): Features, open challenges, and trends,” *Wirel. Commun. Mob. Comput.*, vol. 2018, 2018.

[78] R. Barik, H. Dubey, S. Sasane, C. Misra, N. Constant, and K. Mankodiya, “Fog2Fog: Augmenting Scalability in Fog Computing for Health GIS Systems,” in *IEEE 2nd International Conference on Connected Health: Applications, Systems and Engineering Technologies, CHASE 2017*, 2017, pp. 241–242.

[79] J. Vora, S. Tanwar, S. Tyagi, N. Kumar, and J. J. P. C. Rodrigues,

"FAAL: Fog computing-based patient monitoring system for ambient assisted living," in *IEEE 19th International Conference on e-Health Networking, Applications and Services (Healthcom)*, 2017, pp. 1–6.

[80] I. Azimi *et al.*, "HiCH: Hierarchical Fog-Assisted Computing Architecture for Healthcare IoT," *ACM Trans. Embed. Comput. Syst. Artic.*, vol. 16, no. 20, 2017.

[81] R. H. Dolin *et al.*, "HL7 clinical document architechture, release 2," *J Am Med Inf. Assoc*, vol. 13, pp. 30–39, 2006.

[82] C. Lubamba and A. Bagula, "Cyber-healthcare cloud computing interoperability using the HL7-CDA standard," in *IEEE Symposium on Computers and Communications*, 2017, no. Iscc, pp. 105–110.

[83] H. S. Kim, H. Cho, and I. K. Lee, "The development of a graphical user interface engine for the convenient use of the HL7 version 2.x interface engine," *Healthc. Inform. Res.*, vol. 17, no. 4, pp. 214–223, 2011.

[84] S. K. Sood and I. Mahajan, "A Fog Based Healthcare Framework for Chikungunya," *IEEE Internet Things J.*, vol. 4662, no. c, pp. 1–8, 2017.

[85] Y. Zhao and Y. Zhang, "Comparison of decision tree methods for finding active objects," *Adv. Sp. Res.*, vol. 41, no. 12, pp. 1955–1959, 2008.

[86] S. Aljawarneh, M. B. Yassein, and M. Aljundi, "An enhanced J48 classification algorithm for the anomaly intrusion detection systems," *Cluster Comput.*, vol. 22, no. s5, pp. 10549–10565, 2019.

[87] B. Waters, "Ciphertext-policy attribute-based encryption: An expressive, efficient, and provably secure realization," in *Public Key Cryptohraphy-PKC 2011*, vol. 6571 LNCS, no. subaward 641, 2011, pp. 53–70.

[88] D. Giri, “SecHealth : An Efficient Fog based Sender Initiated Secure Data Transmission of Healthcare Sensors for e-Medical System,” in *EEE Global Communications Conference*, 2017, pp. 1–6.

[89] A. A. Omala, K. P. Kibiwott, and F. Li, “An Efficient Remote Authentication Scheme for Wireless Body Area Network,” *J. Med. Syst.*, vol. 41, no. 2, 2017.

[90] A. Rajagopalan, M. Jagga, A. Kumari, and S. T. Ali, “A DDoS prevention scheme for session resumption SEA architecture in healthcare IoT,” *3rd IEEE Int. Conf.*, pp. 1–5, 2017.

[91] R. Island, “Smart Fog: Fog Computing Framework for Unsupervised Clustering Analytic in Wearable Internet of Things,” in *Global Conference on Signal and Information Processing (GlobalSIP)*, 2017, pp. 472–476.

[92] P. Boersma and V. van Heuven, “Speak and unSpeak with Praat,” *Glot Int.*, vol. 5, no. 9–10, pp. 341–347, 2001.

[93] “Neos guide.” .

[94] S. Samsatli and N. J. Samsatli, “A general mixed integer linear programming model for the design and operation of integrated urban energy systems,” *J. Clean. Prod.*, vol. 191, pp. 458–479, 2018.

[95] Z. Lv, F. Xia, G. Wu, L. Yao, and Z. Chen, “iCare: A mobile health monitoring system for the elderly,” in *IEEE/ACM International Conference on Green Computing and Communications, GreenCom 2010,* 2010.

[96] Cisco, “The Zettabyte Era : Trends and Analysis,” 2017.

[97] R. Prieto, “Cisco Visual Networking Index Predicts Near-Tripling of IP Traffic by 2020,” 2016.

[98] A. Vishwanath, J. Zhu, K. Hinton, R. Ayre, and R. Tucker, “Estimating the Energy Consumption for Packet Processing, Storage and Switching in Optical-IP Routers,” *Opt. Fiber Commun. Conf. Fiber Opt. Eng. Conf. 2013*, p. OM3A.6, 2013.

[99] Alcatel-Lucent, “Alcatel-Lucent 7368 ISAM ONT G-240G-A,” 2014.

[100] R. Unless, P. Act, W. Rose, T. If, and W. Rose, “GreenTouch GreenMeter Core Network Energy Efficiency Improvement Measures and Optimization [Invited],” 2018.

[101] T. E. H. E.-G. and J. M. H. E. A. Q. Lawey, “Distributed energy efficient clouds over core networks,” *J. Light. Technol.*, vol. 32, no. 7, pp. 1261–1281, 2014.

[102] J. Baliga, R. W. A. Ayre, K. Hinton, and R. S. Tucker, “Green Cloud Computing: Balancing Energy in Processing, Storage and Transport,” *Proc. IEEE*, vol. 99, no. 1, pp. 149–167, 2010.

[103] C. Gray, R. Ayre, K. Hinton, and R. S. Tucker, “Power consumption of IoT access network technologies,” in *IEEE International Conference on Communication Workshop (ICCW)*, 2015, pp. 2818–2823.

[104] S. S. D. B. R. H. M. K. J. M. E. H. N. A. I. L. W. Shehabi Arman; Smith, “United States Data Center Energy Usage Report - LBNL-1005775,” no. June, 2017.

[105] A. Q. Lawey, T. E. H. El-Gorashi, and J. M. H. Elmirghani, “Distributed energy efficient clouds over core networks,” *J. Light. Technol.*, vol. 32,

no. 7, pp. 1261–1281, 2014.

[106] A. D. Elgendi M, Eskofier B, Dokos S, “Revisiting QRS Detection Methodologies for Portable , Wearable , Battery-Operated , and Wireless ECG Systems,” *PLoS ONE 9(1) e84018*, vol. 9, no. 1, 2014.

[107] H. Xia, I. Asif, and X. Zhao, “Cloud-ECG for real time ECG monitoring and analysis,” *Comput. Methods Programs Biomed.*, vol. 110, no. 3, pp. 253–259, 2012.

[108] N. Lowres *et al.*, “Identifying postoperative atrial fi brillation in cardiac surgical patients posthospital discharge , using iPhone ECG : a study protocol,” *BMJ Open*, vol. 5:e006849., pp. 1–6, 2015.

[109] N. A. O. Sir Amyas Morse KCB, Comptroller and Auditor General, “Stocktake of access to general practice in England,” 2015.

[110] X. Wang, Q. Gui, B. Liu, Z. Jin, and Y. Chen, “Enabling Smart Personalized Healthcare : A Hybrid Mobile-Cloud Approach for ECG Telemonitoring,” *IEEE J. Biomed. Heal. Informatics*, vol. 18, no. 3, pp. 739–745, 2014.

[111] Public Health England, “Public Health Profiles,” *Public Health England*, 2015. [Online]. Available: http://healthierlives.phe.org.uk/. [Accessed: 20-Jun-2005].

[112] N. Hannent, “Ofcom UK Mobile Sitefinder,” *Ofcom UK Mobile*. [Online]. Available: https://fusiontables.google.com/DataSource?docid=1N4nf1AmXFDk-lbh9Y54jh2FwyudbX3O8-aVIzwZJ#rows:id=1.

[113] Nokia, “LTE-M – Optimizing LTE for the Internet of Things White Paper,”

2015.

[114] “FTTC Exchanges,” *Sam Knows Ltd 2019*, 2016. [Online]. Available: https://availability.samknows.com/broadband/exchanges/bt/fttc.

[115] I. Cale, A. Salihovic, and M. Ivekovic, “GPON (Gigabit Passive Optical Network),” in *29th International Conference on Information Technology Interfaces*, 2007, pp. 679–684.

[116] Eltex, “GPON Optical Line Terminal Data Sheet,” 2015.

[117] Powertec Telecommunications Pty Ltd, “Improving Mobile Signal,” 2014.

[118] A. Lucent, “Leveraging GPON for Mobile Backhaul Networks Overcoming the challenges of mobile broadband adoption,” 2009.

[119] “Cardiovascular Disease Statistic 2015, British Heart Foundation,” *British Cardiovascular Intervention Society*, 2014. .

[120] Offices for National Statistics, “Overview of the UK Population : November 2015,” *Office for National Statistics*, 2015. [Online]. Available: http://www.ons.gov.uk/ons/rel/pop-estimate/overview-of-the-uk-population/index.html.

[121] M. R. Moody GB, “The impact of the MIT-BIH Arrhythmia Database,” *IEEE Eng in Med and Biol 20(3):45-50*. .

[122] A. L. Goldberger *et al.*, “PhysioBank, PhysioToolkit, and PhysioNet Components of a New Research Resource for Complex Physiologic Signals,” *Components a New Res. Resour. Complex Physiol. Signals.*, no. Circulation 101(23):e215-e220 [Circulation Electronic Pages, 2017.

[123] and K. M. H. D. A. M. N. C. M. A. D. B. L. M. Y. S. Q. Y. Umer Akbar,

"Fog Computing in Medical Internet-of-Things: Architecture, Implementation, and Applications," in *Handbook of Large-Scale Distributed Computing in Smart Healthcare*, 2017, pp. 87–119.

[124] A. L. Goldberger, Z. D. Goldberger, and A. Shvilkin, *Goldberger's Clinical Electrocardiography*. 2013.

[125] M. K Turagam, P. Velagapudi, and A. G Kocheril, "Standardization of QRS Duration Measurement and LBBB Criteria in CRT Trials and Clinical Practice," *Curr. Cardiol. Rev.*, vol. 9, no. 1, pp. 20–23, 2013.

[126] Stone, "STONEPC LITE," *Stone Group*, 2017. [Online]. Available: https://www.stonegroup.co.uk/hardware/desktops/lite/.

[127] A. H. El Fawal, A. Mansour, M. Najem, F. Le Roy, and D. Le Jeune, "LTE-M adaptive eNodeB for emergency scenarios," *Int. Conf. Inf. Commun. Technol. Converg. ICT Converg. Technol. Lead. Fourth Ind. Revolution, ICTC*, vol. 2017-Decem, no. August, pp. 536–541, 2017.

[128] A. N. Al-quzweeni, A. Q. Lawey, T. E. H. Elgorashi, and M. H. Jaafar, "Optimized Energy Aware 5G Network Function Virtualization," *IEEE Access*, vol. PP, p. 1, 2020.

[129] T. P. S. Archana, B, "Resource Allocation in LTE : An Extensive Review on Methods , Challenges and Future Scope," *Commun. Appl. Electron.*, vol. 3, no. 2, pp. 12–22, 2015.

[130] R. W. A. A. and R. S. T. A. Vishwanath, F. Jalali, K. Hinton, T. Alpcan, "Energy consumption comparison of interactive cloud-based and local applications," *IEEE J. Sel. Areas Commun.*, vol. 33, no. 4, pp. 616–626, 2015.

[131] V. Valancius, N. Laoutaris, L. Massoulié, C. Diot, and P. Rodriguez, “Greening the internet with nano data centers,” in *5th international conference on Emerging networking experiments and technologies CoNEXT 09*, 2009, p. 37.

[132] G. Auer, O. Blume, V. Giannini, I. Godor, and M. A. Imran, “Energy efficiency analysis of the reference systems, areas of improvements and target breakdown,” *EARTH Proj. Rep. Deliv. D2.3*, pp. 1–68, 2012.

[133] Apple Support, “iMac power consumption and thermal output,” 2017. [Online]. Available: https://support.apple.com/en-gb/HT201918.

[134] T. A. and R. S. T. F. Jalali, K. Hinton, R. Ayre, “Fog Computing May Help to Save Energy in Cloud Computing,” *IEEE J. Sel. Areas Commun.*, vol. 34, no. 5, pp. 1728–1739, 2016.

[135] Cisco, “Cisco Catalyst 6840-X Series Fixed Backbone Switch,” 2016.

[136] U. Gui, “Cisco Industrial Ethernet 2000 Series Switches - Products & Services,” 2016.

[137] Cisco, “Catalyst 6840-X Switch Series Hardware Installation Guide,” 2012.

[138] “Fall and elderly,” *Centers for Disease Contraol and Prevention*, 2016. [Online]. Available: http://www.cdc.gov/media/releases/2016/p0922-older-adult-falls.html.

[139] M. D. R. Jan Gurley , M.D., Nancy Lum , M.S., Merle Sande , M.D., Bernard Lo , M.D., and Mitchell H. Katz, “Person Found in Their Homes Helpless or Dead,” *N. Engl. J. Med.*, vol. 334, no. 26, pp. 1710–1716, 1996.

[140] Age UK, “Later Life in the United Kingdom,” 2015.

[141] F. Emelia J. Benjamin, MD, SCM *et al.*, “Heart Disease and Stroke Statistics,” 2017.

[142] A. S. H. Patricia A. Grady, *Using Nursing Research to Shape Health Policy*. 2017.

[143] Public HealthData Science Team, “Statistic of Patients with Heart Disease,” *Public Health England*, 2018. [Online]. Available: https://fingertips.phe.org.uk/profile-group/cardiovascular-disease-diabetes-kidney- disease/profile/cardiovascular.

[144] G. Mastorakis and D. Makris, “Fall detection system using Kinect ’ s infrared sensor,” *J. Real-Time Image Process.*, vol. 9, no. 4, pp. 635–646, 2014.

[145] A. Inc., “Local and Push Notification Programming Guide,” 2013.

[146] C. Colman-meixner, C. Develder, S. Member, M. Tornatore, S. Member, and B. Mukherjee, “A Survey on Resiliency Techniques in Cloud Computing Infrastructures and Applications,” *IEEE Commun. Surv. Tutorials*, vol. 18, no. 3, pp. 2244–2281, 2016.

[147] R. S. Couto, S. Secci, M. E. M. Campista, and L. H. M. K. Costa, “Server placement with shared backups for disaster-resilient clouds,” *Comput. Networks*, vol. 93, pp. 423–434, 2015.

[148] C. Develder, J. Buysse, B. Dhoedt, and B. Jaumard, “Joint dimensioning of server and network infrastructure for resilient optical grids/clouds,” *IEEE/ACM Trans. Netw.*, vol. 22, no. 5, pp. 1591–1606, 2014.

[149] B. Jan, H. Farman, H. Javed, B. Montrucchio, M. Khan, and S. Ali,

“Energy efficient hierarchical clustering approaches in wireless sensor networks: A survey,” *Wirel. Commun. Mob. Comput.*, vol. 2017, 2017.

[150] I. S. M. Isa, M. O. I. Musa, T. E. H. El-gorashi, A. Q. Lawey, and J. M. H. Elmirghani, “Energy Efficiency of Fog Computing Health Monitoring Applications,” *2018 20th Int. Conf. Transparent Opt. Networks*, pp. 1–5, 2018.

[151] Blueraq Networks Ltd, “Cloud Backup: UK Based, Automatic & Secure,” *BACKUP VAULT*, 2019. .

Appendix 1

Sets, parameters and variables used in the thesis

Sets	
CL	Set of clinics
BS	Set of BSs
ONU	Set of ONUs
OLT	Set of OLTs
CAS	Set of centre aggregation switches
AR	Set of aggregation routers
CR	Set of core routers
CLR	Set of cloud routers
CLS	Set of cloud switches
CS	Set of content servers
CST	Cloud storage
N_m	Set of neighbouring nodes of node m in the network
N	Set of nodes ($N \in CL \cup BS \cup ONU \cup OLT \cup CAS \cup AR \cup CR \cup CLR \cup CLS \cup CS \cup CST$)
FN	Set of candidate locations to deploy PS ($FN \in ONU \cup OLT$ for fog while $FN \in CLS$ for conventional)
Set: Additional for Chapter 4	
ASW	Set of access switches
Set: Additional for Chapter 6	
ND	Set of BSs, ONUs and OLTs (access layer)
Parameters	
$s\ and\ d$	Denote source node s and destination node d of traffic between a node pair

$i\ and\ j$	Denote end nodes of a physical link in the network, $i, j \in N$
P_s	Number of patients in clinic s
IBS	Idle power consumption of a base station (W)
PBS	Power per physical resource block (PRB) of a base station (W/PRB)
Ɍ	Maximum number of PRBs in a base station dedicated for healthcare applications
$PONU$	Maximum power consumption of an ONU (W)
$IONU$	Idle power consumption of an ONU (W)
$CONU$	Maximum capacity of an ONU (bps)
$POLT$	Maximum power consumption of an OLT (W)
$IOLT$	Idle power consumption of an OLT (W)
$COLT$	Maximum capacity of an OLT (bps)
$PCAS$	Maximum power consumption of a centre aggregation switch (W)
$ICAS$	Idle power consumption of a centre aggregation switch (W)
$CCAS$	Maximum capacity of a centre aggregation switch (bps)
PAR	Maximum power consumption of an aggregation router (W)
IAR	Idle power consumption of an aggregation router (W)
CAR	Maximum capacity of an aggregation router (bps)
PCR	Maximum power consumption of a core router (W)
ICR	Idle power consumption of a core router (W)
CCR	Maximum capacity of a core router (W)
$PCLR$	Maximum power consumption of a cloud router (W)
$ICLR$	Idle power consumption of a cloud router (W)
$CCLR$	Maximum capacity of a cloud router (bps)
$PCLS$	Maximum power consumption of a cloud switch (W)

$ICLS$	Idle power consumption of a cloud switch (W)
$CCLS$	Maximum capacity of a cloud switch (bps)
PCS	Maximum power consumption of a content server (W)
ICS	Idle power consumption of a content server (W)
CCS	Maximum capacity of a content server (bps)
$PCST$	Maximum power consumption of a cloud storage (W)
$ICST$	Idle power consumption of a cloud storage (W)
$CCST$	Maximum capacity of a cloud storage (bit)
PPS	Maximum power consumption of a processing server (W)
IPS	Idle power consumption of a processing server (W)
Ωmax	Maximum number of patients per processing server
Λmax	Maximum storage capacity of processing server (bit)
δa	Data rate per patient to send raw health data from clinic to processing server (bps)
τa	Transmission time per patient to send raw health data from clinic to processing server (s)
Ra	Physical resource block per patient to send raw health data from clinic to processing server
α	Size of analysed health data per patient (bit)
δb	Data rate per patient to send analysed health data from processing server to clinic (bps)
τb	Transmission time per patient to send analysed health data from processing server to clinic (s)
Rb	Physical resource block per patient to send analysed health data from processing server to clinic
δc	Data rate per patient to send analysed health data from processing server to cloud storage (bps)

τc	Transmission time per patient to send analysed health data from processing server to cloud storage (s)
δ_{sd}	$\delta_{sd} = 1$ to send the storage traffic from processing servers located at candidate node s, to the cloud storage node d, $s \in FN, d \in CST$
x	Fraction of idle power consumption of networking equipment contributed by the healthcare application under consideration
λ_{ij}	The capacity of link ij dedicated for the healthcare application under consideration (bps)
η	Power usage effectiveness (PUE) of the access network, metro network and IP over WDM network
c	Power usage effectiveness (PUE) of the fog (processing server) and cloud equipment
M	A large enough number
Parameters: Additional for Chapter 4	
$PASW$	Maximum power consumption of an access switch (W)
$IASW$	Idle power consumption of an access switch (W)
$CASW$	Maximum capacity of an access switch (bps)
$PAGS$	Maximum power consumption of an aggregation switch (W)
$IAGS$	Idle power consumption of an aggregation switch (W)
$CAGR$	Maximum capacity of an aggregation switch (bps)
Parameters: Additional for Chapter 5	
N	Maximum number of processing servers per candidate node
Variables	
P_{sd}	Raw health data traffic from source node s to destination node d (bps), $s \in CL, d \in FN$

P_{ij}^{sd}	Raw health data traffic from source node s to destination node d that traverses the link between nodes i and j (bps), $s \in CL, d \in FN,\ i,j \in N$
P_i	Total raw health data traffic that traverses node i (bps), $i \in N$
F_{sd}	Analysed health data feedback traffic from source node s to destination node d (bps), $s \in FN, d \in CL$
F_{ij}^{sd}	Analysed health data feedback traffic from source node s to destination node d that traverses the link between nodes i and j (bps), $s \in FN, d \in CL, i,j \in N$
F_i	Total analysed health data feedback traffic that traverses node i (bps), $i \in N$
S_{sd}	Analysed health data storage traffic from source node s to destination node d (bps), $s \in FN, d \in CST$
S_{ij}^{sd}	Analysed health data storage traffic from source node s to destination node d that traverses the link between nodes i and j (bps), $s \in FN, d \in CST, i,j \in N$
S_i	Total analysed health data storage traffic that traverses node i (bps), $i \in N$
ω_{sd}	Number of patients from clinic s served by processing server located at candidate node d
Pa_{ij}	Number of patients in clinic i served by BS j to send raw health data traffic (integer)
Pb_{ij}	Number of patients in clinic i served by BS j to receive analysed health data feedback traffic (integer)
βa_j	Number of PRBs used in BS j to serve raw health data traffic (integer)
βb_i	Number of PRBs used in BS i to serve analysed health data feedback traffic (integer)

Y_d	$Y_d = 1,$ if a processing server is placed at candidate node $d,$ otherwise $Y_d = 0$, $d \in FN$
ϕ_d	Number of processing servers placed at candidate node d, $d \in FN$
τp_d	Processing and analysis time of processing server (seconds) at candidate node d, $d \in FN$
ζa_j	$\zeta a_j = 1,$ if raw health data traffic traverses node j, otherwise $\zeta a_j = 0$, $j \in N$
ζb_i	$\zeta b_i = 1,$ if analysed health data feedback traffic traverses node i, otherwise $\zeta b_i = 0$, $i \in N$
θc_i	$\theta c_i = 1,$ if analysed health data storage traffic traverses node i where node i is the source of a link, otherwise $\theta c_i = 0$, $i \in N$
ϑc_j	$\vartheta c_j = 1,$ if analysed health data storage traffic traverses node j where j is the end of a link, otherwise $\vartheta c_j = 0$, $j \in N$
ζc_i	$\zeta c_i = 1$, if the analysed health data storage traffic traverses node i where $\zeta c_i = \theta c_i \; OR \; \vartheta c_i$, otherwise $\sigma_i = 0$, $i \in N$
ν_i	ν_i is a dummy variable that takes value of $\theta c_i \oplus \vartheta c_i$, where $\oplus$ is an XOR operation, $i \in N$
EAN	Energy consumption of access network
$ETBS$	Total energy consumption of base stations
$EBSP$	Energy consumption of base stations required to relay raw health data traffic
$EBSF$	Energy consumption of base stations required to relay analysed health data feedback traffic
$ETONU$	Total energy consumption of ONUs

$EONUP$	Energy consumption of ONUs required to relay raw health data traffic
$EONUF$	Energy consumption of ONUs required to relay analysed health data feedback traffic
$EONUS$	Energy consumption of ONUs required to relay analysed health data storage traffic
$ETOLT$	Total energy consumption of OLTs
$EOLTP$	Energy consumption of OLTs required to relay raw health data traffic
$EOLTF$	Energy consumption of OLTs required to relay analysed health data feedback traffic
$EOLTS$	Energy consumption of OLTs required to relay analysed health data storage traffic
EMN	Energy consumption of metro network
$ECASP$	Energy consumption of centre aggregation switches required to relay raw health data traffic
$ECASF$	Energy consumption of centre aggregation switches required to relay analysed health data feedback traffic
$ECASS$	Energy consumption of centre aggregation switches required to relay analysed health data storage traffic
$EARP$	Energy consumption of aggregation routers required to relay raw health data traffic
$EARF$	Energy consumption of aggregation routers required to relay analysed health data feedback traffic
$EARS$	Energy consumption of aggregation routers required to relay analysed health data storage traffic
ECN	Energy consumption of core network
$ECRP$	Energy consumption of core routers required to relay raw health data traffic

$ECRF$	Energy consumption of core routers required to relay analysed health data feedback traffic
$ECRS$	Energy consumption of core routers required to relay analysed health data storage traffic
ECL	Energy consumption of cloud
$ECLRP$	Energy consumption of cloud routers required to relay raw health data traffic
$ECLRF$	Energy consumption of cloud routers required to relay analysed health data feedback traffic
$ECLRS$	Energy consumption of cloud routers required to relay analysed health data storage traffic
$ECLSP$	Energy consumption of cloud switches required to relay raw health data traffic
$ECLSF$	Energy consumption of cloud switches required to relay analysed health data feedback traffic
$ECLSS$	Energy consumption of cloud switches required to relay analysed health data storage traffic
$ECSS$	Energy consumption of content servers required to relay analysed health data storage traffic
$ECSTS$	Energy consumption of cloud storage required to store the analysed health data storage traffic
EFN	Energy consumption of fog nodes
EPS	Energy consumption of processing servers
$ECSN$	Energy consumption of cloud server node
Variables: Additional for Chapter 4	
$ETASW$	Total energy consumption of access switches
$EASWP$	Energy consumption of access switches required to relay raw health data traffic

$EASWF$	Energy consumption of access switches required to relay analysed health data feedback traffic
$EASWS$	Energy consumption of access switches required to relay analysed health data storage traffic
$ETASG$	Total energy consumption of aggregation switches
$EASGP$	Energy consumption of aggregations switches required to relay raw health data traffic
$EAGSF$	Energy consumption of aggregation switches required to relay analysed health data feedback traffic
$EAGSS$	Energy consumption of aggregation switches required to relay analysed health data storage traffic
Variables: Additional for Chapter 6	
ωa_{sd}	Number of patients from clinic s served by primary processing servers located at candidate node d, $s \in CL, d \in FN$
ωb_{sd}	Number of patients from clinic s served by secondary processing servers located at candidate node d, $s \in CL, d \in FN$
Ya_d	$Ya_d = 1$, if one or more primary processing servers are located at candidate node d, otherwise $Ya_d = 0$, $d \in FN$
Yb_d	$Yb_d = 1$, if one or more secondary processing servers are placed at candidate node d, otherwise $Yb_d = 0$, $d \in FN$
z_d	z_d is a dummy variable that takes a value of $Ya_d \oplus Yb_d$, where $\oplus$ is an XOR operation, $d \in FN$
ϕa_d	Number of primary processing servers placed at candidate node d, $d \in FN$
ϕb_d	Number of secondary processing servers placed at candidate node d, $d \in FN$
τpa_d	Processing and analysis time of primary processing server (seconds) at candidate node d, $d \in FN$

τpb_d	Processing and analysis time of secondary processing server (seconds) at candidate node d, $d \in FN$
Pa_{sd}	Raw health data traffic from clinic s to primary processing servers at destination node d (bps), $s \in CL, d \in FN$
Pb_{sd}	Raw health data traffic from source node s to secondary processing servers at destination node d (bps), $s \in CL, d \in FN$
Pa_{ij}^{sd}	Raw health data traffic from source node s to primary processing servers at destination node d that traverses the link between nodes i and j (bps), $s \in CL, d \in FN,\ i, j \in N$
Pb_{ij}^{sd}	Raw health data traffic from source node s to secondary processing servers at destination node d that traverses the link between nodes i and j (bps), $s \in CL, d \in FN,\ i, j \in N$
Fa_{sd}	Analysed health data feedback traffic from primary processing servers at source node s to clinic at node d (bps), $s \in FN, d \in CL$
Fb_{sd}	Analysed health data feedback traffic from secondary processing servers at source node s to clinic at node d (bps), $s \in FN, d \in CL$
Fa_{ij}^{sd}	Analysed health data feedback traffic from primary processing servers at source node s to clinic at node d that traverses the link between nodes i and j (bps), $s \in FN, d \in CL, i, j \in N$
Fb_{ij}^{sd}	Analysed health data feedback traffic from secondary processing servers at source node s to clinic at node d that traverses the link between nodes i and j (bps), $s \in FN, d \in CL, i, j \in N$
Sa_{sd}	Analysed health data storage traffic from primary processing servers at source node s to cloud storage at node d (bps), $s \in FN, d \in CST$
Sb_{sd}	Analysed health data storage traffic from secondary processing servers at source node s to cloud storage at node d (bps), $s \in FN, d \in CST$

Sa_{ij}^{sd}	Analysed health data storage traffic from primary processing servers at source node s to cloud storage at node d that traverses the link between nodes i and j (bps), $s \in FN, d \in CST, i, j \in N$
Sb_{ij}^{sd}	Analysed health data storage traffic from secondary processing servers at source node s to cloud storage at node d that traverses the link between nodes i and j (bps), $s \in FN, d \in CST, i, j \in N$
La_{ij}	$La_{ij} = 1$, if the incoming and/or outgoing traffic of primary processing servers traverses the link between nodes i and j otherwise $La_{ij} = 0$
Lb_{ij}	$Lb_{ij} = 1$, if the incoming and/or outgoing traffic of secondary processing servers traverses the link between nodes i and j otherwise $Lb_{ij} = 0$
ρa_i	$\rho a_i = 1$, if the incoming and/or outgoing traffic of primary processing servers traverse node i, otherwise $\rho a_i = 0$
ρb_i	$\rho b_i = 1$, if the incoming and/or outgoing traffic of secondary processing servers traverses node i, otherwise $\rho b_i = 0$

Appendix 2

Source code for Energy Efficient Fog Computing (EEFC) model

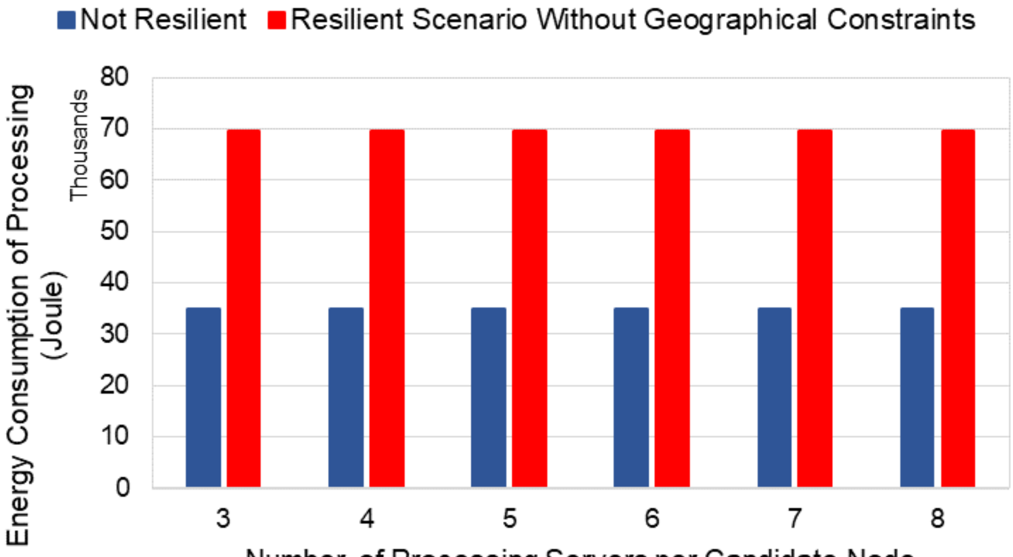
```
#SETS
set N; set CL; set BS; set ONU; set OLT; set CAS; set AR; set CR;
set CLR; set CLS; set CS;
set CST; set FN; set Nm{N} within N;

#PARAMETERS
param Pt{i in CL} >=0;
param PBS >=0; param IBS >=0; param R >=0;
param PONU >=0; param IONU >=0; param CONU >=0;
param POLT >=0; param IOLT >=0; param COLT >=0;
param PCAS >=0; param ICAS >=0; param CCAS >=0;
param PAR >=0; param IAR >=0; param CAR >=0;
param PCR >=0; param ICR >=0; param CCR >=0;
param PCLR >=0; param ICLR >=0; param CCLR >=0;
param PCLS >=0; param ICLS >=0; param CCLS >=0;
param PCS >=0; param ICS >=0; param CCS >=0;
param PCST >=0; param ICST >=0; param CCST >=0;
param PPS >=0; param IPS >=0; param Lambda_max >=0;
param delta_a >=0; param tau_a>=0; param R_a >=0;
param delta_b >=0; param tau_b >=0; param R_b >=0;
param delta_c >=0; param tau_c >=0;
param delta {s in FN, d in CST} >=0; param alpha >=0;
param lambda_a{i in N, j in N} >=0;
param lambda_b{i in N, j in N} >=0;
param lambda_c{i in N, j in N} >=0;
param eta >=0; param c >=0; param M >=0;
param x >=0; param Omega_max >=0; param m_ps >= 0; param c_ps >=0;

#VARIABLES
var Psd{s in CL, d in FN} >=0;
var Psd_ij{s in CL, d in FN, i in N, j in N} >=0;
var P{i in N} >=0;

var Fsd{s in FN, d in CL} >=0;
var Fsd_ij{s in FN, d in CL, i in N, j in N} >=0;
var F{i in N} >=0;
var Ssd{s in N, d in CST} >=0;
var Ssd_ij{s in FN, d in CST, i in N, j in N} >=0;
var S{i in N} >=0;


var omega {s in CL, d in FN} integer >= 0;
var Pa{i in CL, j in BS} integer >=0; var Pb{i in BS, j in CL}
integer >=0;
var beta_a{j in BS} integer >=0; var beta_b{i in BS} integer >=0;
var Y{ i in FN} binary >=0; var phi{i in FN} integer >=0;
var tau_p{i in FN} >=0;
var zeta_a{i in N} binary >=0; var zeta_b{i in N} binary >=0; var
zeta_c{i in N} binary >=0;
var theta_c{i in N} binary >=0; var vartheta_c{i in N} binary >=0;
var v{i in N} binary >=0;


#ENERGY OF ACCESS NETWORK
var EBSP = sum{i in BS}(IBS * x * zeta_a[i] + PBS * beta_a[i]) *
tau_a;
var EBSF = sum{i in BS}(IBS * x * zeta_b[i] + PBS * beta_b[i]) *
tau_b;
var ETBS = EBSP + EBSF;


var EONUP = sum{i in ONU}(IONU * x * zeta_a[i] + P[i] * ((PONU-
IONU)/CONU) ) * tau_a;
var EONUF = sum{i in ONU}(IONU * x * zeta_b[i] + F[i] * ((PONU-
IONU)/CONU) ) * tau_b;
var EONUS = sum{i in ONU}(IONU * x * zeta_c[i] + S[i] * ((PONU-
IONU)/CONU) ) * tau_c;
var ETONU = EONUP + EONUF + EONUS;


var EOLTP = sum{i in OLT}(IOLT * x * zeta_a[i] + P[i] *((POLT-
IOLT)/COLT) ) * tau_a;
var EOLTF = sum{i in OLT}(IOLT * x * zeta_b[i] + F[i] *((POLT-
IOLT)/COLT) ) * tau_b;
var EOLTS = sum{i in OLT}(IOLT * x * zeta_c[i] + S[i] *((POLT-
IOLT)/COLT) ) * tau_c;
var ETOLT = EOLTP + EOLTF + EOLTS;


var EAN = (ETBS + ETONU + ETOLT) * eta;


#ENERGY OF METRO NETWORK
var ECASS = sum{i in CAS}(ICAS *x * zeta_c[i] + S[i] * ((PCAS-
ICAS)/CCAS) ) * tau_c;
var EARS = sum{i in AR}(IAR * x * zeta_c[i] + S[i] * ((PAR-IAR)/CAR)
) *tau_c;
```

```
var EMN = (ECASS +EARS) * eta;


#ENERGY OF CORE NETWORK
var ECRS = sum{i in CR}(ICR * x * zeta_c[i] + S[i] * ((PCR-ICR)/CCR)
) * tau_c;
var ECN = ECRS * eta;


#ENERGY OF CLOUD
var ECLRS = sum{i in CLR}(ICLR * x * zeta_c[i] + S[i] * ((PCLR-
ICLR)/CCLR) ) * tau_c;
var ECLSS = 2 * sum{i in CLS}(ICLS * x * zeta_c[i] + ((PCLS -
ICLS)/CCLS) ) * tau_c;
var ECSS = sum{i in CS}(ICS * x * zeta_c[i] + S[i] * ((PCS-ICS)/CCS)
) * tau_c;
var ECSTS = 2 * sum{i in CST}(ICST * x * zeta_c[i] + S[i] * tau_c *
((PCST - ICST)/CCST) ) * tau_c;
var ECL = (ECLRS +ECLSS + ECSS + ECSTS) * c;


#ENERGY OF FOG NODE
var EPS = sum{i in FN}(IPS * phi[i] * (tau_a + tau_b + tau_c) + PPS
* tau_p[i]);
var EFN = EPS * c;


#OBJECTIVE
minimize energy : EAN + EMN + ECN + ECL + EFN;


#CONTRAINS
#ASSOCIATION PATIENTS TO A PROCESSING SERVER
s.t. A1 {s in CL, d in FN} : omega[s,d] <= Pt[s] * Y[d];
s.t. A2 {s in CL} : sum{d in FN} omega[s,d] = Pt[s];


#TRAFFIC FROM CLINICS TO PROCESSING SERVER
s.t. B1 {s in CL, d in FN} : Psd[s,d] = omega[s,d] * delta_a;


#TRAFFIC FROM PROCESSING SERVER TO CLINICS
s.t. C1 {s in FN, d in CL} : Fsd[s,d] = omega[d,s] * delta_b;


#TRAFFIC FROM PROCESSING SERVER TO CLOUD STORAGE
s.t. D1 {s in FN, d in CST} : Ssd[s,d] = sum{i in CL} omega[i,s] *
delta_c * delta[s,d];


#FLOW CONSERVATION
s.t. E1 {s in CL, d in FN, i in N} : sum{j in
Nm[i]:i<>j}Psd_ij[s,d,i,j] - sum{j in Nm[i]:i<>j}Psd_ij[s,d,j,i] =
if i=s then Psd[s,d] else if i=d then -Psd[s,d]  else 0;


s.t. E2 {s in FN, d in CL, i in N} : sum{j in
Nm[i]:i<>j}Fsd_ij[s,d,i,j] - sum{j in Nm[i]:i<>j}Fsd_ij[s,d,j,i] =
if i=s then Fsd[s,d] else if i=d then -Fsd[s,d]  else 0;


s.t. E3 {s in FN, d in CST, i in N} : sum{j in
Nm[i]:i<>j}Ssd_ij[s,d,i,j] - sum{j in Nm[i]:i<>j}Ssd_ij[s,d,j,i] =
if i=s then Ssd[s,d] else if i=d then -Ssd[s,d]  else 0;


#TOTAL TRAFFIC TRAVERSING NODE
s.t. F1 {i in N} : P[i] = sum{s in CL,d in FN, j in
Nm[i]:s<>d&&i<>j}Psd_ij[s,d,j,i];
s.t. F2 {i in N} : F[i] = sum{s in FN,d in CL, j in
Nm[i]:s<>d&&i<>j}Fsd_ij[s,d,i,j];
s.t. F3 {i in N} : S[i] = sum{s in FN,d in CST, j in
Nm[i]:s<>d&&i<>j}Ssd_ij[s,d,j,i] + sum{d in CST:i<>d}Ssd[i,d];


#LINK CAPACITY CONSTRAINT
s.t. G1 {i in N, j in Nm[i]:i<>j} : sum{s in CL, d in FN}
Psd_ij[s,d,i,j] <= lambda_a[i,j]; #link from BS to Clinic is 0
s.t. G2 {i in N, j in Nm[i]:i<>j} : sum{s in FN, d in CL}
Fsd_ij[s,d,i,j] <= lambda_b[i,j]; #link from Clinic to BS is 0
s.t. G3 {i in N, j in Nm[i]:i<>j} : sum{s in FN, d in CST}
Ssd_ij[s,d,i,j] <= lambda_c[i,j]; #link from Clinic to BS is 0


#NODE TO TRANSMIT RAW HEALTH DATA TO PS
s.t. H1 {j in N} : sum{s in CL, d in FN, i in
N:i<>j}Psd_ij[s,d,i,j]*100000 >= zeta_a[j];
s.t. H2 {j in N} : sum{s in CL, d in FN, i in
N:i<>j}Psd_ij[s,d,i,j]*100000 <= M * zeta_a[j];


#NODE TO TRANSMIT ANALYSED HEALTH DATA TO CLINIC
s.t. I1 {i in N} : sum{s in FN, d in CL, j in
Nm[i]:i<>j}Fsd_ij[s,d,i,j] *100000 >= zeta_b[i];
s.t. I2 {i in N} : sum{s in FN, d in CL, j in
Nm[i]:i<>j}Fsd_ij[s,d,i,j] *100000 <= M * zeta_b[i];


#NODE TO TRANSMIT ANALYSED HEALTH DATA TO CLOUD
s.t. J1 {i in N} : sum{s in FN, d in CST, j in
Nm[i]:i<>j}Ssd_ij[s,d,i,j] *100000 >= theta_c[i];
s.t. J2 {i in N} : sum{s in FN, d in CST, j in
Nm[i]:i<>j}Ssd_ij[s,d,i,j] *100000 <= M * theta_c[i];
s.t. J3 {j in N} : sum{s in FN, d in CST, i in
N:i<>j}Ssd_ij[s,d,i,j] *100000 >= vartheta_c[j];
```

```
s.t. J4 {j in N} : sum{s in FN, d in CST, i in
N:i<>j}Ssd_ij[s,d,i,j] *100000 <= M * vartheta_c[j];
s.t. J5 {i in N} : theta_c[i] + vartheta_c[i] = 2 * zeta_c[i] -
v[i];


#NUMBER OF PRB TO SEND RAW HEALTH DATA
s.t. K1 {i in CL, j in BS:i<>j} : Pa[i,j] = sum{s in CL, d in FN
:s<>d} (Psd_ij[s,d,i,j]/delta_a);
s.t. K2 {i in CL} : sum{j in BS} Pa[i,j] = Pt[i];
s.t. K3 {j in BS} : beta_a[j] = sum{i in CL} Pa[i,j] * R_a;
s.t. K4 {j in BS} : beta_a[j] <= R;


#NUMBER OF PRB TO SEND ANALYSED HEALTH DATA
s.t. L1 {i in BS, j in CL} : Pb[i,j] = sum{s in FN, d in
CL:s<>d}(Fsd_ij[s,d,i,j]/delta_b);
s.t. L2 {j in CL} : sum{i in BS}Pb[i,j] = Pt[j];
s.t. L3 {i in BS} : beta_b[i] = sum{j in CL}Pb[i,j] * R_b;
s.t. L4 {i in BS} : beta_b[i] <= R;


#MAX PATIENTS PER PS
s.t. M1 {d in FN} : sum{s in CL} omega[s,d] <= Omega_max * phi[d];


#PROCESSING AND ANALYSIS TIME AT EACH PS
s.t. P1 {d in FN} : tau_p[d] = sum{s in CL}m_ps * omega[s,d] + c_ps
*phi[d];


#MAX STORAGE CAPACITY PER PS
s.t. Q1 {d in FN} : sum{s in CL}omega[s,d] * alpha <= Lambda_max *
phi[d];
```

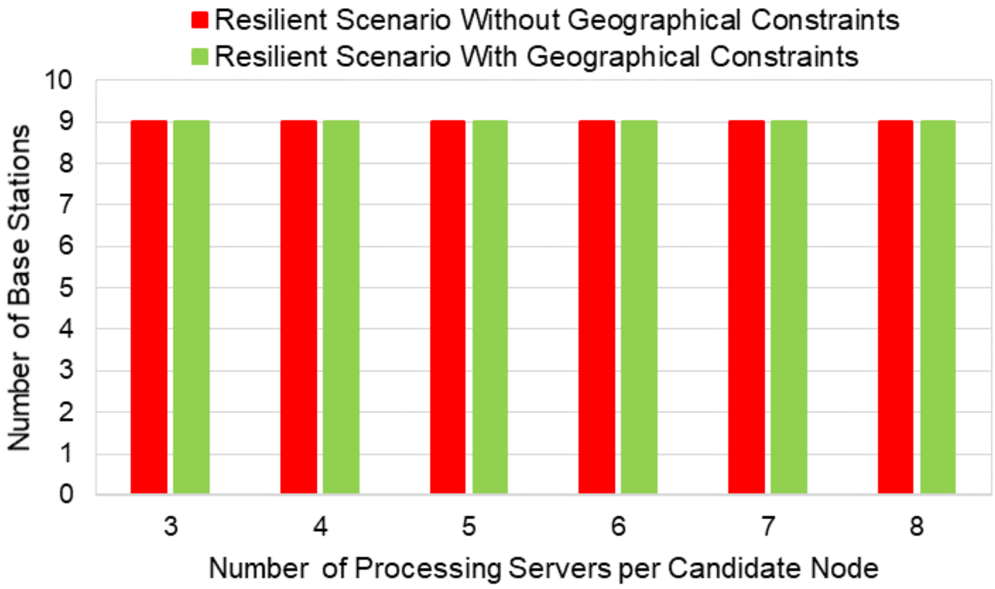
Appendix 3

Energy consumption of processing for non-resilient scenario and resilient scenario without geographical constraints, for fall monitoring applications

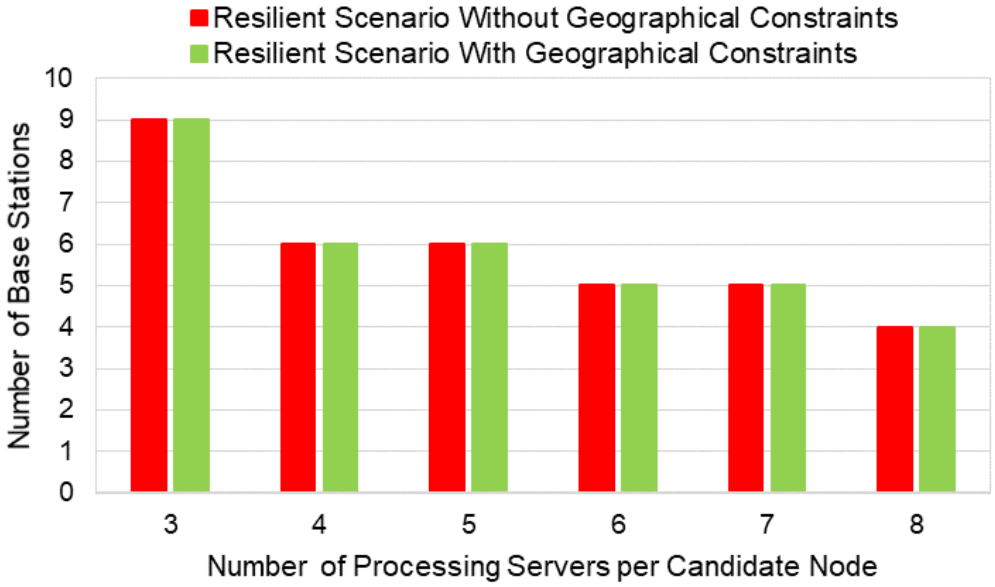


Appendix 4

Number of base stations used to send (a) the raw video signal for processing and (b) the analysed video signal for feedback, for a resilient scenario, without geographical constraints and resilient scenario, with geographical constraints under different number of processing servers per candidate node for fall monitoring applications.



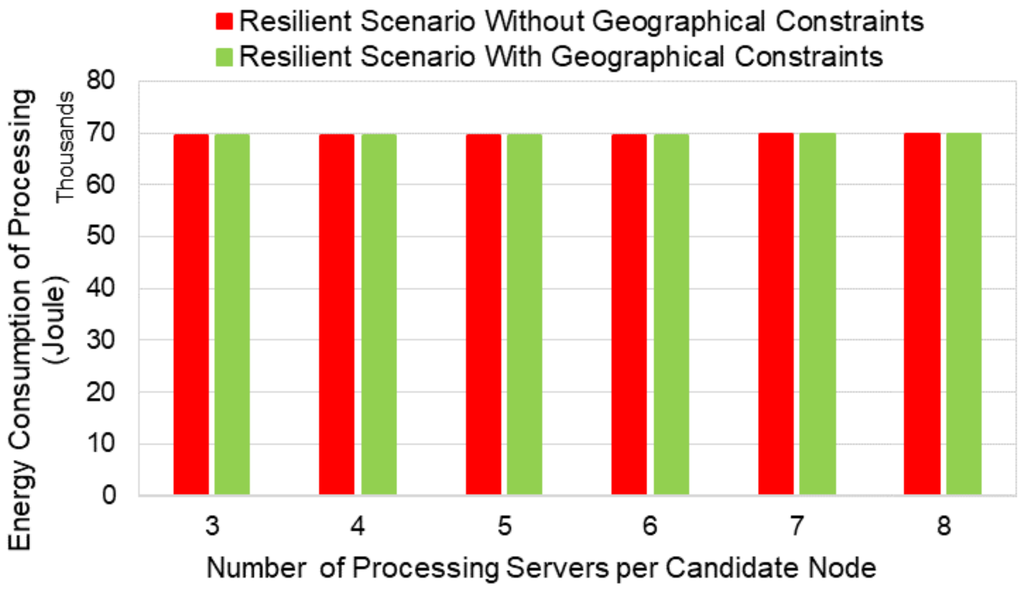
(a)



(b)

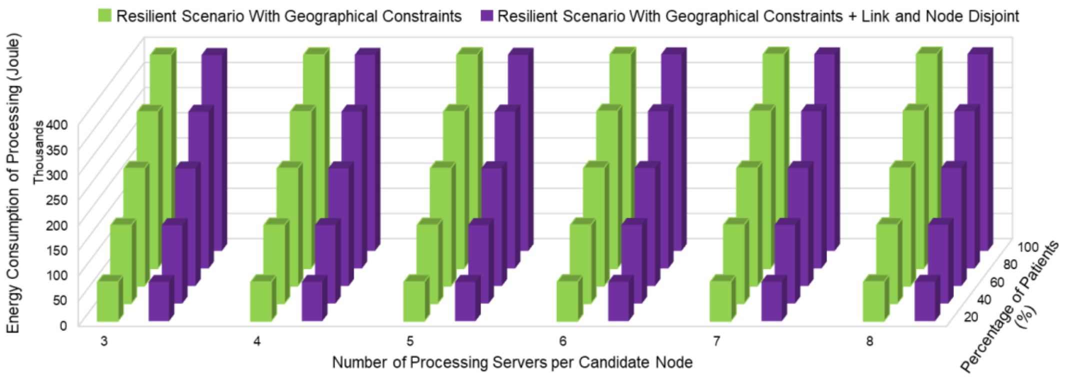
Appendix 5

Energy consumption of processing for the resilient scenario without geographical constraints and the resilient scenario, considering geographical constraints for fall monitoring applications



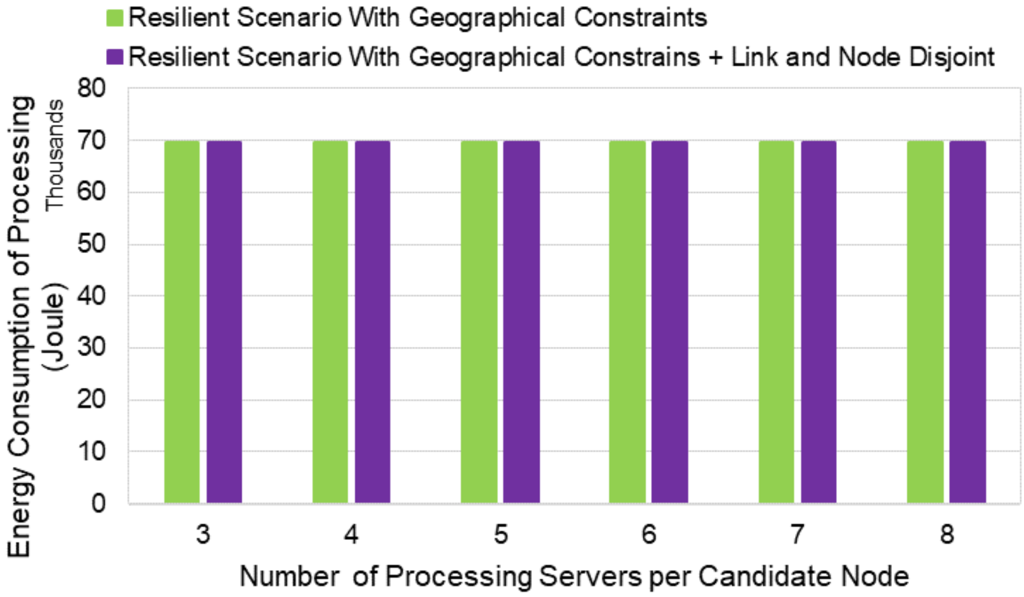
Appendix 6

Energy consumption of processing for the resilient scenario considering the geographical constraints; and the energy consumption of the resilient scenario with geographical constraints and link and node disjoint resilience for ECG monitoring applications



Appendix 7

Energy consumption of processing for the resilient scenario considering the geographical constraints and the resilient scenario considering the geographical constraints and link and node disjoint resilience for fall monitoring applications


🟩 Resilient Scenario With Geographical Constraints
🟪 Resilient Scenario With Geographical Constrains + Link and Node Disjoint
Energy Consumption of Processing (Joule)
Thousands
Number of Processing Servers per Candidate Node