ENHANCED K-MEANS CLUSTERING WITH NOMA, SWIPT AND HARQ, FOR MASSIVE M2M COMMUNICATION

by

Emerson Cabrera

for the degree of Doctor of Philosophy

School of Engineering
Macquarie University

February 15, 2020

Supervisor: Rein Vesilo

Contents

ABSTRACT

One of the operating modes of the 5G cellular network, termed **Massive Machine-to-Machine Communication (MMC)**, is currently limited by the scarcity of network resources in Long Term Evolution-Advanced (LTE-A) and the shortage of device battery power. MMC is expected to be utilised by applications such as sensor nodes, where attempted simultaneous access of network resources through the Random Access CHannel (RACH) in LTE-A, has been shown to lead to an *overload and access problem*. Therefore, this thesis proposes an enhanced K-means Clustering algorithm accompanied by Non-Orthogonal Multiple Access (NOMA), to enable the MMC operating mode. 'User Pairing' is applied to each cluster, with the strongest channel device assigned as the cluster head (CH), to enhance the network sum throughput. Energy Harvesting (EH) through Simultaneous Wireless Power and Information Transfer (SWIPT) is incorporated to address the DL Successive Interference Cancellation (SIC) and UL transmission power consumption concerns, and to increase the energy-efficiency (EE). A performance analysis was conducted, where our proposed scheme was shown: (**a**) to have a higher network sum throughput than the traditional K-means with a minimum rate requirement of 100–2000 kbps; and (**b**) that SIC and UL data transmission with SWIPT is feasible, with a minimum rate requirement of 100–1700 kbps.

Another operating mode of 5G, termed **Ultra-Reliable Communication (URC)**, is expected to be utilised by applications such as mission critical industrial control, medical and Vehicle-to-Everything (V2X). These applications will be under the constraints of high availability, ultra-reliability and ultra-low latency. Hybrid Automatic Repeat reQuest (HARQ) has been proven to improve the reliability of data transmission by the retransmission (RTX) of erroneous packets during poor channel conditions. This thesis also proposes an enhanced NOMA HARQ scheme to accompany the enhanced K-means clustering algorithm, to improve the RTX process and therefore reduce the delay incurred from the possibility of multiple RTX. A performance analysis was conducted, by comparing

our enhanced NOMA HARQ scheme with the competing Chase Combining (CC) HARQ and the LTE-A HARQ OMA scheme, in terms of outage probability, reliability, and delay. Our enhanced NOMA HARQ scheme was shown to have a lower outage probability and higher achievable rate compared to CC-HARQ and LTE-A HARQ OMA scheme. This was due to the incorporation of Incremental Redundancy (IR) HARQ and cooperative relaying NOMA. A lower outage probability led to increased reliability and therefore decreased delay compared to the CC-HARQ and LTE-HARQ OMA schemes.

Naturally, these two operating modes of 5G are closely linked together because 5G is regarded as an important enabler for the Internet of Things (IoT). The IoT consists of Massive Machine-Type Communication (mMTC) and Ultra-Reliable Low-Latency Communication (URLLC), which are the massive form of Machine-to-Machine (M2M) Communication and URC with low-latency communication, respectively. Therefore, M2M and URC were both addressed in this thesis, by extending the proposed enhanced K-means Clustering algorithm NOMA scheme with SWIPT, to include an enhanced HARQ NOMA scheme.

STATEMENT OF CANDIDATE

I, Emerson Cabrera, declare that this report, submitted as part of the requirement for the award of Doctor of Philosophy in the School of Engineering, Macquarie University, is entirely my own work unless otherwise referenced or acknowledged. This document has not been submitted for qualification or assessment at any academic institution.

Candidate's Signature:

Date: September 13, 2019

ACKNOWLEDGMENTS

I would like to acknowledge:

- **Dr Rein Vesilo** for sharing his immense knowledge and giving excellent feedback on my research work as my supervisor and mentor. He provided motivation and guidance during my time spent researching and writing this thesis.
- **Dr Gengfa Fang** for his collaborative role for the research contained in our VTC2017-Spring conference paper on an enhanced HARQ scheme called 'A-HARQ'.
- **Xunqian Tong** for allowing me to be a co-author for his journal paper, which is related to my field of research on 5G cellular networks, specifically on the use of clustering for the proposed enhanced K-means clustering with NOMA scheme.

I would also like to thank my girlfriend for her continued support and understanding, and also my parents and younger sister for their endless support and encouragement.

Publications and Awards

1. E. Cabrera and R. Vesilo, "**Enhanced K-means Clustering and Uplink Non-Orthogonal Multiple Access (NOMA) for Massive M2M Communication**," *submitted to IEEE Internet of Things (IoT) Journal*, 2019. [1] contributes to the research contained in Chapter 4.

2. E. Cabrera and R. Vesilo, "**An Enhanced K-means Algorithm with Non-Orthogonal Multiple Access (NOMA) for MMC Networks**," in *2018 28th International Telecommunication Networks and Applications Conference (ITNAC)*, November 2018, pp. 1-8. [2] contributes to the research contained in Chapter 3.

3. E. Cabrera, G. Fang and R. Vesilo, "**Adaptive Hybrid ARQ (A-HARQ) for Ultra-Reliable Communication in 5G**," in *2017 IEEE 85th Vehicular Technology Conference (VTC Spring)*, June 2017, pp. 1-6. [3][1] contributes to the research contained in Chapter 5.

4. X. Tong, G. Fang, D. Nguyen, J. Lin and E. Cabrera, "**An Energy-Balanced Routing Algorithm in Wireless Seismic Sensor Network**", *Journal of Computational and Theoretical Nanoscience*, vol. 13, no. 10, pp. 6823-6833, October 2016. [5] contributes to the literature review contained in Chapter 2.

5. Awarded the '**Australian Postgraduate Award (APA)**', to supplement the research contained in this thesis.

[1] the results in [3] were partly included in Chapter 4 of my MRes thesis in 2015 [4].

List of Figures

List of Tables

List of Acronyms

3GPP	3rd Generation Partnership Project
4G/5G	4th Generation/5th Generation cellular network
AWGN	Additive White Gaussian Noise
CC	Chase Combining
CDMA	Code Division Multiple Access
CSI	Channel State Information
D2D	Device-to-Device
DL	Downlink
EE	Energy-Efficiency
EH	Energy Harvesting
ER	Energy Rich
eMBB	Enhanced Mobile Broadband
eNodeB	LTE-A Evolved NodeB
FDD	Frequency Division Duplexing
FDMA	Frequency Division Multiple Access
H2H	Human-to-Human
HARQ	Hybrid Automatic Repeat reQuest
IMT	International Mobile Telecommunications
IoT	Internet of Things
IR	Incremental Redundancy
LPWA	Low Power Wide Area
LTE/LTE-A	Long Term Evolution/LTE-Advanced
M2M/MMC	Machine-to-Machine/Massive M2M Communication
MA	Multiple Access
MCS	Modulation and Coding Scheme

MIMO	Multiple Input Multiple Output
MRC	Maximal Ratio Combining
MTC/mMTC	Machine-Type Communication/Massive MTC
NOMA	Non-Orthogonal Multiple Access
NR	5G New Radio
OFDM	Orthogonal Frequency Division Multiplexing
OFDMA	Orthogonal Frequency Division Multiple Access
OMA	Orthogonal Multiple Access
QoS	Quality of Service
RACH	Random Access CHannel
RAT	Radio Access Technology
RB	Resource Block
RFID	Radio Frequency ID
RTX	Retransmission
Rx	Receiver
SIC	Successive Interference Cancellation
SINR/SNR	Signal-to-Interference-and-Noise Ratio/Signal-to-Noise Ratio
SWIPT	Simultaneous Wireless Power and Information Transfer
TDD	Time Division Duplexing
TDMA	Time Division Multiple Access
TTI	Transmission Time Interval
Tx	Transmitter
UE	User Equipment
UL	Uplink
URC	Ultra Reliable Communication
URLLC	Ultra-Reliable Low-Latency Communication
V2X	Vehicle-to-Everything

Chapter 1

Introduction

Cellular networks[1] have evolved from generation to generation at roughly one decade apart, where each successive generation has offered progressively higher data rates, and also additional features for cellular network subscribers.

The 1st Generation (1G) was rolled out in 1981, involving cellular communication with analogue transmissions using circuit switching, which only allowed voice calls. The 2nd Generation (2G) was rolled out in 1992, with the addition of Short Message Service (SMS) text messages, Electronic Mail (Email), and also allowed global roaming. The 3rd Generation (3G) was rolled out in 2001, with the addition of mobile internet. 2G used the Multiple Access (MA) techniques of Time Division MA (TDMA) and Frequency Division MA (FDMA), while 3G used Code Division MA (CDMA) [6].

Prior to the roll out of the 4th Generation (4G)[2] in 2011, LTE (technically 3.9G[3]) was rolled out in 2008, with the addition of Multiple Input Multiple Output (MIMO), and optimised support for packet-switched data services; which was standardised in the 3rd Generation Partnership Project (3GPP) Release 8. LTE-A was then rolled out in 2011, as an enhancement over LTE, with the addition of wider bandwidths (through Carrier Aggregation), heterogeneous networks, relaying, and Coordinated MultiPoint (CoMP) transmission and reception; which was standardised in 3GPP Release 10 (LTE-A is backwards compatible with LTE). 4G used the MA techniques of Orthogonal Frequency-Division MA (OFDMA), TDMA and FDMA [6].

Later additions to 4G is LTE-A Pro (technically 4.5G) in 2016, with the additions of

[1] parts of this page was previously published in **Chapter 1** of my MRes thesis in 2015 [4].

[2] also known as Long Term Evolution-Advanced (LTE-A).

[3] since it does not completely adhere to the IMT-Advanced standard.

Low Power Wide Area (LWPA) cellular technologies to support the Internet of Things (IoT), cellular Vehicle-to-Everything (V2X), etc; which was standardised in 3GPP Release 13. 3GPP Releases 14–16+, has/will provide further improvements to LTE-A Pro, on the road towards the 5th Generation (5G), as shown in Figure 1.1 [7].

By 2022, it is predicted that the IoT will feature approximately 14.6 billion Internet Protocol (IP) connected devices (3.9 billion on mobile); leading to a significant 7-fold increase in Machine-to-Machine (M2M) IP and mobile data volume (from 2017 levels), and accounting for 51% (31% on mobile) of global connections [8, 9]. By comparison, LTE-A was mainly designed for a *relatively* small number of human operated devices exchanging large packets, where the feedback and control signalling is negligible; rather than for the IoT, where the overhead is significant [10, 11]; and was not guaranteed to provide high reliability for the majority of the time [12]. Therefore, 5G has been widely considered as an important enabler for the IoT [11, 13–15].

This chapter details the 5G cellular network, the requirements and challenges for 5G IoT (including mMTC and URLLC), the potential solutions to overcome these challenges, and also the thesis specifics. Section 1.1 covers 5G, Section 1.2 covers the requirements of 5G IoT, Section 1.3 covers the current challenges for 5G IoT and the potential solutions to overcome them, and Section 1.4 establishes the scope, objectives, contributions, and the outline of this thesis.

1.1 5th Generation New Radio (5G NR)

5G New Radio (NR) was first standardised as '5G Phase 1' in 3GPP Release 15 in 2018, but will be *fully* rolled out in the 2020s[4]. Standardisation of 5G was split into two phases [17], with '5G Phase 2' expected to be completed by June 2020 [18]. 3GPP Release 17 will further enhance 5G, which is expected to be completed by 2021 [19]. 5G is an enhancement over 4G, and will natively support the three major operating modes of **Enhanced Mobile Broadband** (eMBB), **Massive Machine-to-Machine Communication** (MMC), and **Ultra-Reliable Communication** (URC)[5] [10, 21–23]. These operating modes will provide massive capacity, massive connectivity, and ultra-reliable

[4] As of 2019, commercial implementation of eMBB has commenced in countries such as the USA, China, Australia, South Korea, etc [7, 16].

[5] MMC and URC are also known as **Massive Machine-Type Communication** (mMTC) and **Ultra-Reliable Low-Latency Communication** (URLLC), respectively [20].

Figure removed due to copyright restrictions

Figure 1.1: 3GPP Release 13–16+ for LTE-A Pro and 5G NR. © 2018 Qualcomm Technologies, Inc., from [7].

low-latency communications, for a wide range of services, devices, deployments, environments and mobility states. The background for 5G is covered as follows, with heterogeneity and potential applications in Subsection 1.1.1 and the operating regions/modes in Subsection 1.1.2.

1.1.1 Heterogeneity and Potential Applications

As described in the Horizon 2020 FANTASTIC-5G project [24], 5G is expected to be heavily heterogeneous [13], which can be accommodated by using network slicing [25, 26]:

1. **Diverse set of services**: eMBB, MMC/mMTC, URC with ultra-low latency (i.e. URLLC), mission critical industrial control applications, medical applications, Vehicle-to-Everything (V2X), etc [11, 27].
2. **Diverse set of devices**: low-end sensors to high-end tablets.
3. **Different deployments**: macro-cell, femto-cell, small-cell, etc.
4. **Different environments**: low-density, ultra-dense urban, etc.
5. **Different mobility states**: static, low and high (i.e. up to 500 km/h [27]).

There are many potential applications that could benefit from 5G [21, 27], such as (**i**) e-banking, e-health, and e-learning; (**ii**) augmented reality (AR) and virtual reality (VR)

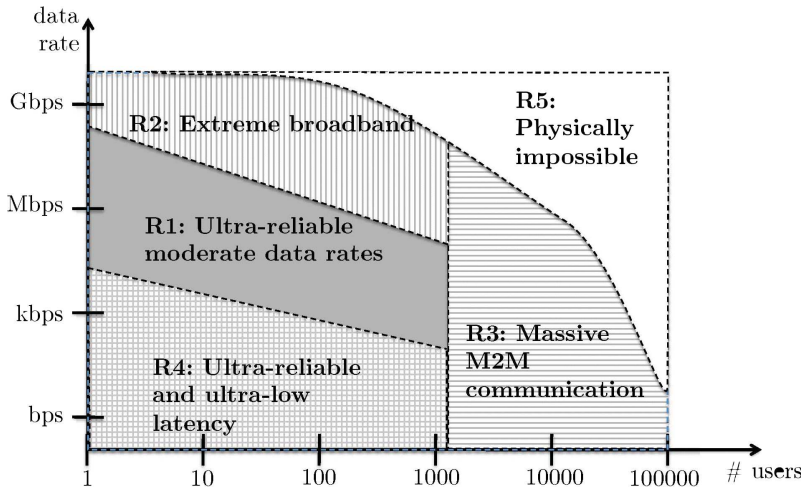


Figure 1.2: Operating regions of 5G cellular networks. © 2014 IEEE, from [28].

offices; (**iii**) sensors and actuator networks; (**iv**) smart-grid network; (**v**) traffic control systems; (**vi**) security, logistics, automotive and mission critical industrial control applications, and; (**vii**) crowded areas such as events and shopping centres.

Some of these applications require ultra-reliability, but others such as mission critical industrial control applications also require ultra-low latency. EE is important for limited power devices such as UE. This makes it very relevant for protocols being able to support the highly diverse sets of applications in the future.

1.1.2 Operating Regions/Modes

There are five expected operating regions (as shown in Figure 1.2) based on the data rate and the number of users in order of magnitude [28]. These operating regions are summarised as follows:

- **R1**: the URC operating mode with moderate data rates, for services such as mobile cloud or Tactile Internet.
- **R2**: the eMBB operating mode will utilise the frequency bands in Frequency Range (FR): FR1 (i.e. sub-6 GHz) and FR2 (i.e. above 24 GHz) [29–31], and technologies such as Massive Multiple Input Multiple Output (mMIMO), full-duplex communication, etc.
- **R3**: the MMC operating mode will utilise frequency bands in FR1 [29–31], and has the additional challenge of supporting redundant or low importance messages (e.g.

sensor alarms for the smart grid) and also high importance messages (e.g. mission critical industrial control applications).

- **R4**: the URC operating mode with ultra-low latency (i.e. URLLC). URLLC will utilise frequency bands in FR1 [29–31], and devices will transmit very short messages (i.e. 32 Bytes in $\leq$ 1 ms, within an urban-macro environment, of which 12 Bytes is protocol overhead [32]) at low data rates. Services such as V2X occupy this region.

Based on the scope of this thesis, regions **R3** and **R4** are the most relevant, with the proposed MMC and URLLC solutions covered in Chapters 3–5.

1.2 Internet of Things (IoT)

LPWA cellular technologies of LTE Cat-M1/Enhanced MTC (eMTC), Extended Coverage-Global System for Mobile Communications for the IoT (EC-GSM-IoT), and LTE Cat-NB1/Narrow-band IoT (NB-IoT), were introduced in 3GPP Release 13 (as shown in Figure 1.1 on page 3), for the IoT [15].

The advantage that LPWA cellular technologies have over the propriety technologies of SigFox, LoRa, and Ingenu Random Phase MA (RPMA), is that cellular technologies operate over licensed spectrum. These licensed LTE-A and GSM spectrum offer high security, interference-free communication and QoS guarantees. They are also scalable, depending on the IoT use case, while the propriety technologies would require new Base Stations to cover future increases in device numbers [15, 33]. However, the LPWA propriety technologies are already deployed in various markets, meaning SigFox and LoRa are currently more cost-effective solutions despite the performance limitations (due to operating over unlicensed spectrum) [34].

1.2.1 Requirements

The following requirements for an efficient deployment of the IoT (consisting of mMTC and URLLC) in 5G [14, 15, 35], are summarised below:

1. Low device cost and low deployment cost.

2. High data rates for high-definition content such as video streaming, VR, AR, etc.

3. High scalability to support fine-grained fronthaul network decomposition through Network Function Virtualization (NFV).

4. Sub-ms latency for applications such as tactile Internet, AR, autonomous cars, drones, eHealth, industrial control, etc.

5. Improved coverage and handover efficiency for devices, especially for smart metering and in elevators.

6. Security of user connectivity and privacy for future applications such as mobile payments and digital wallets over the IoT. However, users must still be traceable for situations such as criminal investigations.

7. Low energy consumption and high EE for low-power and low-cost devices.

8. Prompt delivery of messages within a certain timeframe and area for a massive number of devices.

9. Device-to-Device (D2D) connectivity for a massive number of high mobility devices and for billions of LPWA IoT devices.

There are many applications that would benefit from a successful deployment of the IoT in 5G. These are smart/secure homes, intelligent transportation, smart cities, industrial control, eHealth, drones, etc [15, 35–37].

1.2.2 Problems to be solved for MMC and URLLC

The problems to be solved in relation to MMC and URLLC are: (**i**) the scheduler in the MAC layer for LTE-A treats all packets equally regardless of their content, rather than giving higher priority to more important packets; (**ii**) LTE-A cannot guarantee high reliability for the majority of the time, as the current network architecture is not designed to provide this level of service [12]; (**iii**) EE is not optimal enough for the energy limited UE. This can be a major problem for MMC/IoT, as these devices need to be able to survive for years on limited amounts of power, and; (**iv**) latency can increase to unacceptable levels during poor channel conditions, as the probability of ≥ 1 RTX is high, which can cause problems for URLLC applications.

Most of these problems are in relation to mission critical industrial control and medical applications, but are also problems for other applications (such as sensor devices), as there is always a balance required between reliability, latency and EE.

1.3 5G IoT Challenges

There are many challenges to be overcome on the road towards 5G, but this thesis will instead focus on enabling the MMC operating mode, which has differing Quality of Service (QoS) requirements, such as massive connectivity, small packets, low data rate, high energy-efficiency (EE), and low energy consumption [11,36,38]. Sensor nodes are used as motivations for the proposed solution [1,2], as the IoT will mostly consist of these devices.

M2M Communication involves the transfer of data between typically non-human operated battery limited devices [39,40], while MMC consists of billions of these M2M devices exchanging small packets with each other [10,21,41]. Non-network enabled devices can be indirectly connected to the network by using Radio Frequency IDs (RFID), by a network enabled device using a RFID reader. This is the basic framework for the IoT [2,42].

These MMC devices could attempt to simultaneously access network resources through the Random Access CHannel (RACH) in LTE-A [41]. Since there are only 54 contention-based preambles available (every 5 ms) during the Random Access (RA) procedure (covered in Subsubsection 2.3.1.1 on page 21), then there is a high possibility of an *overload and access problem* [36,43,44], leading to a significant waste of resources and energy. This possibility can be reduced by placing the devices into clusters [43,45–48] based on criteria, such as shortest distance, channel gain, or Signal-to-Noise Ratio (SNR) [2].

The *overload and access problem* can be avoided by applying the Non-Orthogonal Multiple Access (NOMA) technique, since it does not require the RA procedure [49]. NOMA commonly achieves MA via the code- or power-domain [36,49–55], where this thesis will focus on power-domain NOMA. In power-domain NOMA, signals are superimposed at the transmitter (Tx), and can be separated while mitigating the interference from other devices by the Successive Interference Cancellation (SIC) process at the receiver (Rx).

However, the SIC process has high complexity within an MMC network [49], leading to significant power consumption. This complexity can be reduced (and therefore the power consumption) by applying the 'User Pairing' concept (i.e. pairing strong channel gain devices with weak channel gain devices) [54], which is similar to clustering. The

power consumption can be mitigated by using the Simultaneous Wireless Information and Power Transfer (SWIPT) energy harvesting (EH) technique, which involves splitting the received RF signal or the transmission time slot for EH and data transmission [56–61].

Since the IoT consists of MMC and URLLC devices [14, 15, 35], then this thesis will also focus on addressing the additional QoS requirements of ultra-reliability and ultra-low latency. One proven way of improving reliability in LTE-A is by using Hybrid Automatic Repeat reQuest (HARQ), combining error correction with packet recovery at the physical (PHY) and medium access control (MAC) layers [3, 62, 63]. Type-I HARQ involves the retransmission (RTX) of the erroneous packet, while Type-II HARQ modifies the Modulation and Coding Scheme (MCS) for each RTX, involving differing Redundancy Versions (RV) — containing different combinations of systematic and parity bits [3, 64, 65].

1.3.1 Clustering

Collision resolution is required if two or more devices within a network were to select the same preamble during the RACH RA procedure between the device and eNodeB. Collision resolution solves the issue related to the high possibility of an *overload and access problem* within the IoT/MMC, but leads to increased latency as the collided devices are directed to compete in a future RA Opportunity (RAO) within the time domain [66, 67].

Therefore, it may be better to reduce the impact of MMC devices that can simultaneously attempt to access resources during the RACH RA procedure. This is because reserving RBs in future RAO for collision resolution is inefficient and a waste of resources if there are no collisions during that particular point in time.

Clustering is proposed to solve some challenges faced by MMC (as given later in Subsection 2.2.4 on page 20). Some of these problems are the amount of control signalling overhead, the collisions within the RACH, how to simultaneously accommodate periodic and non-periodic IoT traffic, and the EE of devices (some devices require at least 10 years of battery power [33]).

Clustering can reduce the control signalling overhead [68], which is now more significant due to the expected short length of messages transferred. This is done by arranging the devices within a network into clusters based on certain criteria (i.e. shortest distance, channel gain, SNR, etc), and delegating the role of cluster head (CH) to one of the devices within each cluster. The CH is responsible for communication between the eNodeB and its own cluster, thereby reducing the control signalling to just a single set for the whole

cluster [68]. Clustering also reduces the possibility of an *overload and access problem*, therefore reducing the significant waste of resources and energy.

Since the devices within an MMC network are arranged into clusters, then devices can communicate with each other using D2D Communication, thereby increasing the reliability of transmission within a cluster, and also allowing more RACH RA procedure preambles to be available for contention [68]. This would lead to a reduction of collisions within the RACH, and would also reduce the amount of delay and energy used compared to if a collision had occurred. However, clustering only reduces the possibility of an *overload and access problem*, it does not prevent it.

1.3.2 Non-Orthogonal Multiple Access (NOMA)

Multiple Access (MA) techniques in cellular networks can be categorised into Orthogonal MA (OMA) and NOMA. Examples of OMA techniques are TDMA, FDMA, OFDMA and Single-Carrier FDMA (SC-FDMA), which are used in 4G and earlier generations of cellular networks. Resources are divided among devices in an orthogonal manner, to avoid intra-cell interference and to establish a simple air-interface design. However there is no way to mitigate inter-cell interference, without careful cell planning and interference mitigation techniques [69].

OMA has been shown to be inadequate in fulfilling the high demands of 5G [70]. Power-domain NOMA is widely considered an enabling MA technique for 5G over the conventional OMA [51, 54, 71–73], as it can support the massive connectivity requirement for MMC. NOMA also offers a better trade-off between system throughput and user fairness compared to OMA, because strong channel gain users can utilise the same resources as weak channel gain users, which is otherwise inaccessible in OMA [52, 55, 71, 72, 74]; and it does not require significant modifications to the current LTE-A architecture, unlike with code-domain NOMA [51]. This is why this thesis will focus on power-domain NOMA, as stated in Section 1.3. The *overload and access problem* can be avoided by using NOMA, since it does not require the RA procedure [49].

There are some challenges [51, 69, 74–77] that must be considered if 5G is to utilise NOMA as its primary MA technique. These are:

1. The eNodeB needs to estimate the arrival rate of devices to be able to perform SIC in uncoordinated scenarios (i.e. devices use random slotted access).

2. Devices must be able to perform channel state information (CSI) estimation in uncoordinated scenarios.

3. The SIC process is not error free [78]. The signal of the n-th device, $x_n(t)$, must be regenerated to be able to subtract it from the superimposed signal, $y_n(t)$.

4. Inter-cell interference can be a problem for dense networks.

5. The larger a NOMA cluster is, the more likely SIC process errors and inter-cell interference will affect the network sum throughput.

6. Cell-edge users would require a strong channel gain user to be able to perform SIC perfectly.

7. The complexity of SIC is a significant problem in large un-clustered networks for battery-limited devices.

1.4 Thesis Specifics

1.4.1 Scope

The IoT will feature a combination of mMTC and URLLC devices in 5G. Mission critical industrial control and medical applications, have strict QoS requirements such as ultra-reliability and ultra-low latency (i.e. URLLC), while MMC networks have additional requirements of massive connectivity, small packets, low data rate, high EE and low energy consumption, etc. These IoT devices will most likely communicate over the widely deployed cellular networks.

Based on the challenges faced by the MMC and URLLC operating modes, the subject areas of clustering, NOMA, SWIPT and HARQ are within the scope of this thesis. Within those subject areas, other wireless technologies (such as D2D, WiFi, etc), MIMO NOMA, and green energy SWIPT are out of the scope of this thesis. To simulate the proposed solutions, only the Additive White Gaussian Channel (AWGN) model with the Rayleigh fading channel model is within the scope. Other channel modes are out of the scope of this thesis.

1.4.2 Objectives

Consequently, the objectives (**O1**–**O4**) of this thesis are:

O1. Propose a clustering algorithm with the NOMA technique, to account for the scenario where MMC devices could simultaneously attempt to access network resources, leading to the *overload and access problem*.

O2. Enhance the proposed clustering algorithm by taking into account 'user pairing' NOMA [54] (covered in Subsection 2.5.6 on page 40), for an enhanced NOMA network sum throughput.

O3. Mitigate the power consumption (and therefore increase the EE) of the NOMA DL SIC process and UL data transmission for battery-limited devices, by incorporating the SWIPT EH technique.

O4. Address the ultra-reliability and ultra-low latency QoS requirements for URLLC devices, by using HARQ to ensure the reliability of data transmission and therefore reduce delay.

1.4.3 Contributions

1. An enhanced K-means clustering algorithm accompanied with DL NOMA and SWIPT, is proposed in Chapter 3.

    - By fulfilling the objectives **O1** and **O3**, the MMC operating mode was shown to be feasible while also addressing the power consumption of the NOMA DL SIC process for battery-limited devices, for cluster sizes of larger than 2 (with the SWIPT EH technique in Step 4 of **Algorithm 3.3**). Different from [58], **Algorithm 3.3** considers multiple clusters for SWIPT within an example 50-device network, compared to only a 2-device network.
    - The proposed algorithm experienced a 34.94% higher throughput (25 run average) over OMA, and an improvement of 13.47% over the traditional K-means and 22.00% over Hierarchical, at a minimum rate requirement of 100 kbps.
    - The traditional and improved K-means algorithms (i.e. **Algorithm 2.2** and **Algorithm 3.1**) clusters devices based on their similarity to other devices.

Therefore, the NOMA network sum throughput gain over OMA is not necessarily maximised due to the similarity in device channel gains. Different from **Algorithm 2.2** and **Algorithm 3.1**, the proposed **Algorithm 3.2** utilises 'user pairing', to ensure each cluster has a strong channel gain device as their CH, and any clusters of size 1 formed (i.e. no cluster NOMA throughput gain over OMA) were eliminated. The optimal number of clusters was also determined by using the 'Silhouette Value' metric and the 'Sum of Squared Error' metric, prior to the utilisation of 'user pairing'.

2. An enhanced K-means clustering algorithm accompanied with UL NOMA and SWIPT, is also proposed in Chapter 4.

   - By fulfilling the objectives **O1** and **O3**, the MMC operating mode was shown to be feasible while also addressing the power consumption of UL data transmission for battery-limited devices, for cluster sizes of larger than 2 (with the SWIPT EH technique in Step 4 of **Algorithm 4.2**).
   - The proposed algorithm experienced a 24.77% higher throughput (25 run average) over OMA, and an improvement of 24.06% over the traditional K-means and 24.49% over Hierarchical, at a minimum rate requirement of 100 kbps.
   - By fulfilling objective **O2**, the proposed **Algorithm 4.1** has been enhanced as in the DL. However, different from **Algorithm 3.2**, there was an extra step of re-allocating the second strongest channel gain devices in each 4-device cluster to the closest 2-device cluster, since the traditional and improved K-means algorithms were shown to have a tendency of making inappropriately sized clusters for the UL.
   - The proposed algorithm was further enhanced to simultaneously accommodate DL NOMA and UL NOMA, with SWIPT incorporated to facilitate DL SIC and UL transmission.

3. An enhanced NOMA HARQ scheme for the URC operating mode with ultra-low latency (i.e. URLLC), is proposed in Chapter 5.

   - By addressing objective **O4**, the URLLC operating mode was shown to be feasible while also addressing the power consumption of NOMA DL SIC and of UL transmission for battery-limited devices, even if RTXs are required.

- Different from the proposed NOMA HARQ scheme in [79], the proposed **Algorithm 5.2** included RVs of the UE1 and UE2 signals, which could then be transmitted by the eNodeB at full power instead of split power, for some of the SIC decoding scenarios. This enabled a higher SINR, achievable rate and therefore better outage probability compared to the proposed NOMA HARQ Scheme in [79].

1.4.4 Outline of Later Chapters

Chapter 2 conducts a literature review on MTC in cellular networks, the IoT (including MMC and URLLC), clustering algorithms, NOMA, SWIPT and HARQ. Chapter 3 presents the proposed DL enhanced K-means clustering solution with NOMA and SWIPT, while Chapter 4 presents the proposed UL enhanced K-means clustering solution and the combined DL/UL clustering solution, with NOMA and SWIPT. Chapter 5 presents the proposed enhanced NOMA HARQ scheme to ensure URC and ultra-low latency (i.e. URLLC). Chapter 6 concludes this thesis, and establishes a pathway towards planned future research.

Chapter 2

Literature Review

2.1 Introduction

Wireless cellular networks are widely deployed worldwide, and also offers high security, interference-free communication, QoS guarantees, and are also scalable depending on the IoT use case, as mentioned in Section 1.2. However, the possible simultaneous access of network resources by using the LTE-A RACH RA procedure, by the MMC devices within the IoT, has high potential to be a bottleneck, will eventually lead to the *overload and access problem*. As mentioned in Sections 1.3 and 1.3.1, clustering can significantly reduce the feedback and control signalling overhead (which is more significant for the IoT), and the effects of the *overload and access problem*. However, clustering does not prevent the *overload and access problem* if the OMA technique is used, and can still lead to excessive delays (due to the communication between the CH and eNodeB) and a waste of resources and energy. Devices also must rely on the CHs to reliably transmit their data to the eNodeB. As mentioned in Sections 1.3 and 1.3.2, the NOMA technique can avoid the *overload and access problem*, but clustering is still required due to the high complexity of the NOMA SIC process (and therefore high power consumption) in un-clustered networks. As mentioned in Section 1.3, the power consumption from the DL SIC process or UL data transmission, can be mitigated by using the SWIPT EH technique. The IoT also consists of URLLC devices with additional QoS requirements of ultra-reliability and ultra-low latency. These requirements can be addressed by using HARQ, which has been proven to improve the reliability of data transmission.

Therefore, this chapter conducts a literature review on the related work of MTC in

cellular networks (LTE-A, 5G NR, frequency and time resources, and IoT challenges) in Section 2.2, the IoT (including MMC and URLLC) in Section 2.3, clustering algorithms in Section 2.4, NOMA in Section 2.5, SWIPT in Section 2.6 and HARQ in Section 2.7. Section 2.8 concludes this chapter.

2.2 Machine-Type Communication (MTC) in Cellular Networks

2.2.1 Long Term Evolution-Advanced (LTE-A)

The major differences between LTE and LTE-A[1], where enhancements over LTE were introduced in 3GPP Release 10, are summarised below [6]:

- LTE-A supports up to 100 MHz bandwidths, through carrier aggregation of five Component Carriers of the 20 MHz maximum in LTE. This enables backward compatibility with LTE.
- An enhanced DL multiple antenna transmission is maintained by the number of antennas at the eNodeB and UE, which increases the number of antenna ports from 4 to 8 in Single-User Multiple Input Multiple Output (SU-MIMO).
- Uplink multiple antenna transmission is achieved by increasing the number of antenna ports from 1 to 4 in SU-MIMO, which increases the peak spectral efficiency to 15 bps/Hz and also introduces transmit diversity for UL control signalling.
- Relaying for parts of the network where wired backhaul is impractical. A relay node transfers the control information and data, to and from a donor cell.
- Support for heterogeneous network deployments, which consists of a layer of macro-cells and a layer of small-cells with at least one common carrier between them. Cross-carrier scheduling is used to avoid control channel interference between the macro-cells and small-cells, when control signalling is transferred between them.

[1] parts of Sections 2.2.1, 2.2.3 and 2.2.3.1, were previously published in **Chapter 1** of my MRes thesis in 2015 [4].

2.2.2 5th Generation New Radio (5G NR)

There are three major design objectives (**D1**–**D3**) for 5G, as proposed in [27] and [30]: (**D1.**) support for massive capacity (10 Tbps/km^2) and massive connectivity (millions of devices/km^2); (**D2.**) support for the increasingly diverse set of services, applications and users, and; (**D3.**) flexible and efficient use of all available non-contiguous spectrum.

The following specifications (**S1**–**S5**) can be partially satisfied if these proposed design objectives are successfully implemented [21, 27]: (**S1.**) at least 1 Gbps for ultra high definition (UHD) video and 10 Gbps for mobile cloud services; (**S2.**) ultra-wide bandwidth, as in the 'mmWave', of the 30–300 GHz extremely high frequency (EHF) range [80], and sub-ms latency; (**S3.**) high reliability, in relation to QoS requirements for mission critical industrial control applications, V2X communication, IoT applications, smart sensors, text-based messaging for a smart city [42], etc; (**S4.**) 'Zero-second switching' (i.e. up to 10 ms) between different Radio Access Technologies (RATs), to enable 'Ubiquitous' communications [24], and; (**S5.**) 100-fold lower energy consumption, compared to LTE-A [81].

2.2.3 Time and Frequency Resources

LTE-A resources are split into the time, frequency and spatial domains. The spatial dimension can be accessed by the multiple antenna ports at the Evolved NodeB (eNodeB), where each antenna port has a reference signal (RS) that enables the user equipment (UE) to estimate the radio channel.

In the time domain, resources are split into 10 ms frames, as shown in Figure 2.1. Each frame consists of ten 1 ms subframes, with each subframe consisting of two 0.5 ms slots. Each slot consists of 7 Orthogonal Frequency Division Multiplexing (OFDM) symbols of 66.667 μs, all preceded by a 4.69 μs cyclic prefix (CP), except for the first OFDM symbol with a 5.2 μs CP instead. The CPs are used to prevent inter-symbol interference (ISI), which are due to the varying lengths of the several transmission paths [62].

In the frequency domain, the total bandwidth is split into a number of RBs as shown in **Table 2.1**, which is the minimum resource unit that can be allocated to a UE. Each RB consists of 12 subcarriers of 15 kHz spacing.

The most common mode of operation, where the downlink (DL) and uplink (UL) can simultaneously operate at differing carrier frequencies (separated by a guard band

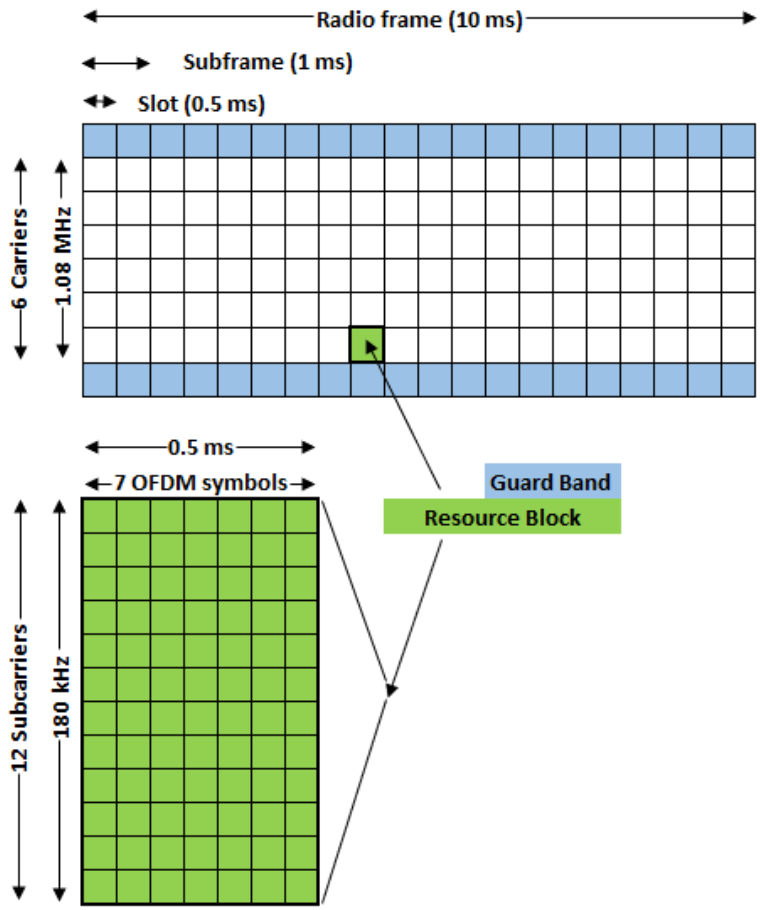


Figure 2.1: The division of resources within the time and frequency domains for an example 1.4 MHz bandwidth in LTE-A.

to prevent interference from a neighbouring channel), is referred to as Frequency Division Duplexing (FDD). The other mode of operation, where the DL and UL have non-simultaneous access to the same carrier frequency (separated by a guard period when switching from a DL subframe to UL), is referred to as Time Division Duplexing (TDD) [6].

2.2.3.1 Resource Scheduling and Interference Reduction

The resource scheduler of LTE-A operates at the MAC layer of the IP stack. UEs queue for the available resources, which are in the form of RBs. The MAC Scheduler allocates RBs to UEs waiting in the queue, which means that it applies equal treatment regardless of the content of the particular packet [6].

Such 'fair treatment' of incoming packets can be a problem for mission critical industrial control applications, as they require ultra-reliability and ultra-low latency. 4G

Table 2.1: Bandwidth resource allocation [6]. **NOTE: 'Bandwidth Utilisation' refers to the portion of the bandwidth available for data transmission.**

Bandwidth (MHz)	Resource Blocks	Bandwidth Utilisation (%)
1.4	6	77.1
3	15	90
5	25	90
10	50	90
15	75	90
20	100	90

cannot guarantee that these strict requirements will be satisfied, since it was not designed to provide such a level of service for the majority of the time. Therefore, 5G has been proposed as a way to provide ultra-reliability and ultra-low latency for UEs.

2.2.3.2 PHY layer changes from LTE-A to 5G

Some of the relevant PHY layer changes in 5G for MMC and URLLC [30, 31, 82–84], are summarised as follows:

- Additional subcarrier spacings are available in 5G, such as 30, 60, 120 and 240 kHz. The LTE-A TTI of 1 ms can thus be reduced by factors of 2, 4, 8 and 16, to 0.5, 0.25, 0.125 and 0.0625 ms.
- Each OFDM symbol can be configured to be UL, DL or 'flexible' (some examples are shown in Figure 2.2). This is contrary to LTE-A, where all OFDM symbols in a slot are configured to be only UL or only DL.
- Slots can be broken up into 'mini-slots', with OFDM symbol lengths of 2, 4 and 7, for flexible starts and lengths for DL/UL traffic. Mini-slots further shortens the TTI duration. Slots can also be aggregated to increase coverage for eMBB applications and to reduce UL/DL switching overhead.
- Low-Density Parity Check (LDPC) and polar codes are used for the data and control channel respectively. The LDPC codes can be used to generate additional parity bits, similar to Type-II/IR HARQ in LTE-A. LDPC codes can achieve low coding rates with higher coding gains for applications that require ultra-reliability.

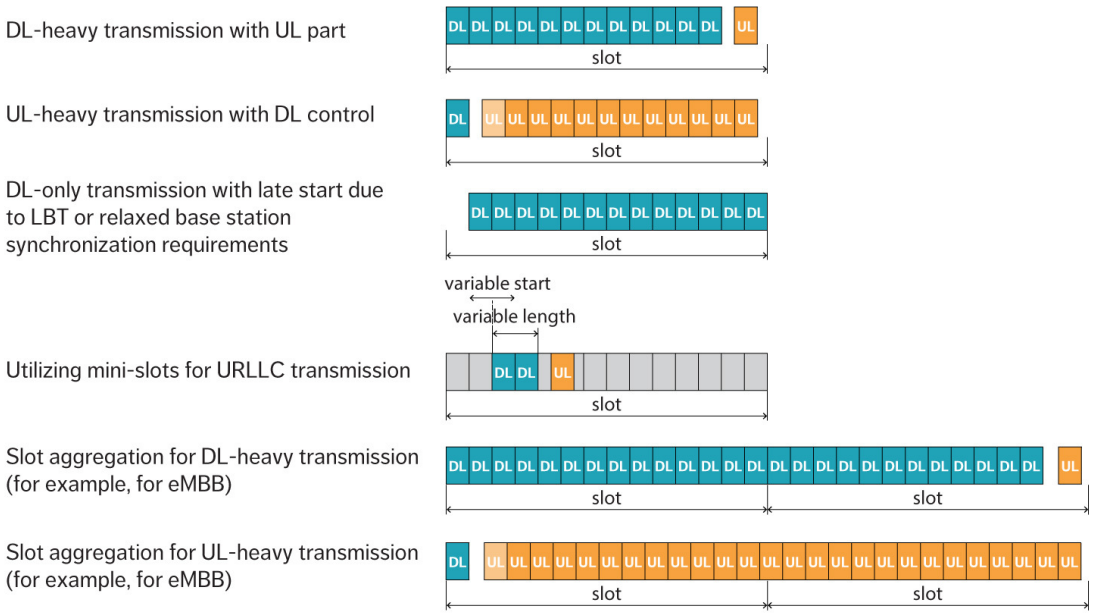


Figure 2.2: Example usage of subframes with 14 OFDM symbols in 5G, where the blank white spaces are 'flexible' symbols. © 2017 Ericsson, from [82].

2.2.4 Internet of Things Challenges

mMTC is a subset of devices within the IoT, where transmitted message lengths are expected to be short. Some of these challenges (**M1**–**M4**) are:

M1. the control overhead for a short message packet will be of a much longer length than the data itself [10].

M2. as mentioned earlier in Section 1.3, an *overload and access problem* is likely to occur (e.g. 99.97% probability, with 1000 devices arriving at 30 ms intervals [85]) when utilising the MMC operating mode, due to the nature of the RACH RA procedure in the current LTE-A architecture [36, 43, 86]. This would lead to many collisions, and therefore a significant waste of resources and energy.

M3. how can periodic and non-periodic traffic be accommodated for the IoT.

M4. some devices are expected to last for at least 10 years on battery power [33, 36], meaning they need to be as energy efficient as possible, since frequently replacing batteries may be impractical.

URLLC enables services such as factory automation, remote control, VR, AR, V2X communications, health care, smart grid, etc [32, 87, 88]. URLLC is another subset of devices within the IoT, which pose additional challenges (**U1**–**U2**) for the IoT:

U1. devices require ≤ 1 ms latency within an urban-macro environment [26, 32], where the following time components must be considered, such as the time to get a transmission grant, Tx signal processing, Rx signal processing, the TTI, and HARQ latency [89]. Processing times would be significantly shorter with grant-free HARQ [90].

U2. devices require messages to be sent with an ultra-reliable probability of $P_{\text{UR}} \geq 99.999\%$ [10, 21, 23, 32].

2.3 Internet of Things

2.3.1 Massive Machine-Type Communication (mMTC)

There are many solutions proposed within the related literature that deal with some of the mMTC challenges (**M1**–**M4**) mentioned in Subsection 2.2.4, with the main examples involving collision resolution, the use of clustering, and resource allocation.

2.3.1.1 RACH RA Procedure

The RACH RA procedure in LTE-A consists of 4 steps (as shown in Figure 2.3) [6, 44]:

1. Devices select one of the 54 available contention-based preambles, out of the available 64 preambles (10 are reserved for devices requiring contention-free access).

2. The eNodeB sends a RA response (RAR) to each device. If there is a collision with another device or a device does not receive a RAR in time, then devices begin again at step 1 after a certain amount of time. Otherwise, the eNodeB directs successful devices on which resource to use for Layer 2/Layer 3 message transmission.

3. Successful devices transmit their Layer 2/Layer 3 message. Note that collisions in step 1 may not be detected, meaning devices that selected the same preamble will collide in step 3.

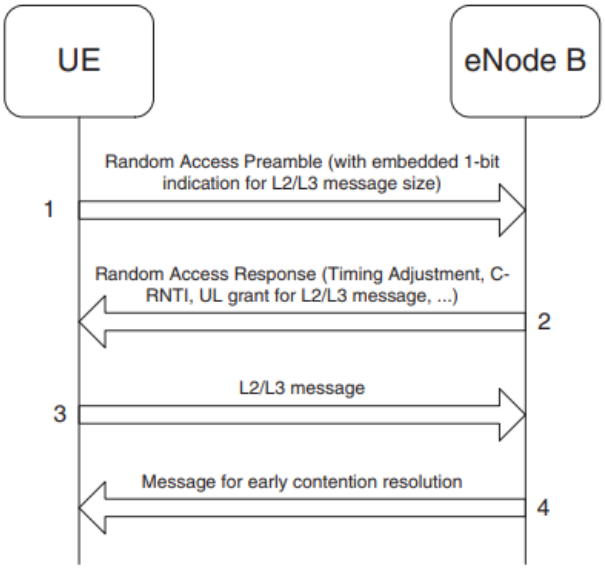


Figure 2.3: The 4-step contention-based RACH RA procedure in LTE-A. © 2011. Reprinted from [6], by permission of John Wiley & Sons Ltd.

4. The eNodeB sends a contention resolution message. Successful devices can transmit using their allocated resource. Unsuccessful devices begin again from step 1.

2.3.1.2 Collision Resolution using Contention-Based Tree-Splitting

The contention-based tree-splitting technique was proposed in [66] and [67], to solve MMC challenge **M2** (i.e. the high possibility of an *overload and access problem*), for when a RACH overload occurs.

In [66], the eNodeB sends a different type of message (MSG 4b) to devices that collided during step 3 of the RA procedure. These devices are directed to select one of the q preambles within a future tree-splitting RAO (TRAO), as shown in Figure 2.4. In [67], a certain number of the 54 contention-based preambles are reserved for MTC, termed 'virtual RA frame'. This number is flexible based on the observed collision rate of MTC devices. The algorithm in [67] differs from [66], because a flexible number of m preambles are reserved in the following virtual RA frame, for all collided devices during the previous TRAO. Both algorithms continue until all collisions have been resolved, or the maximum number of preamble transmissions has been reached.

The significance of [66] and [67], is that when more q/m preambles are available within each TRAO, the average number of preamble transmissions per device is lower.

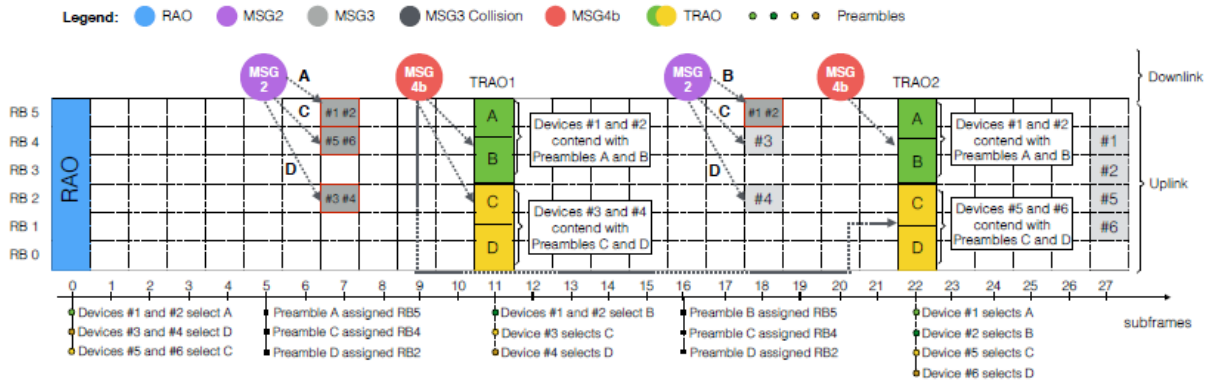


Figure 2.4: Madueno *et al.*'s tree-splitting algorithm for 6 devices and 4 available preambles. © 2014 IEEE, from [66].

The average outage probability in channel access for the algorithm in [67] is also lower. However, this is at the cost of increased delay, because there are less resources available within the next TRAO to resolve any collisions from previous TRAO. The system capacity used to resolve the collisions for the algorithm in [66] is also much lower than the LTE-A dynamic allocation scheme (which has 10 RAOs per frame compared to the usual 2 RAOs per frame [67]).

Evidently, these schemes can resolve the inevitable collisions (if the OMA technique is used) within an MMC/IoT network. However, this is at a cost of a huge 1200 ms delay (at a reliability of 99%), for the best case scenario of 6 q preambles available within each TRAO for 30000 devices. For delay-tolerant IoT applications, this 1200 ms delay or 99% reliability is sufficient to meet their requirements. However, for delay-sensitive URLLC devices, these collision resolution schemes are not ideal, as they violate URLLC challenges **U1** and **U2** (i.e. an ultra-reliable probability of $P_{\text{UR}} \geq 99.999\%$, within a latency of ≤ 1 ms).

2.3.1.3 Clustering based on Quality of Service (QoS) Requirements

Clustering based on QoS requirements was proposed in [39] to also solve challenge **M2**, as multiple M2M devices may have differing QoS requirements. Multiple devices would be segregated based on their QoS requirements, where applications (with deterministic timing constraints) such as medical or industrial control, would have reserved RBs over applications (with statistical timing constraints) such as gaming signals, which would have opportunistically allocated RBs.

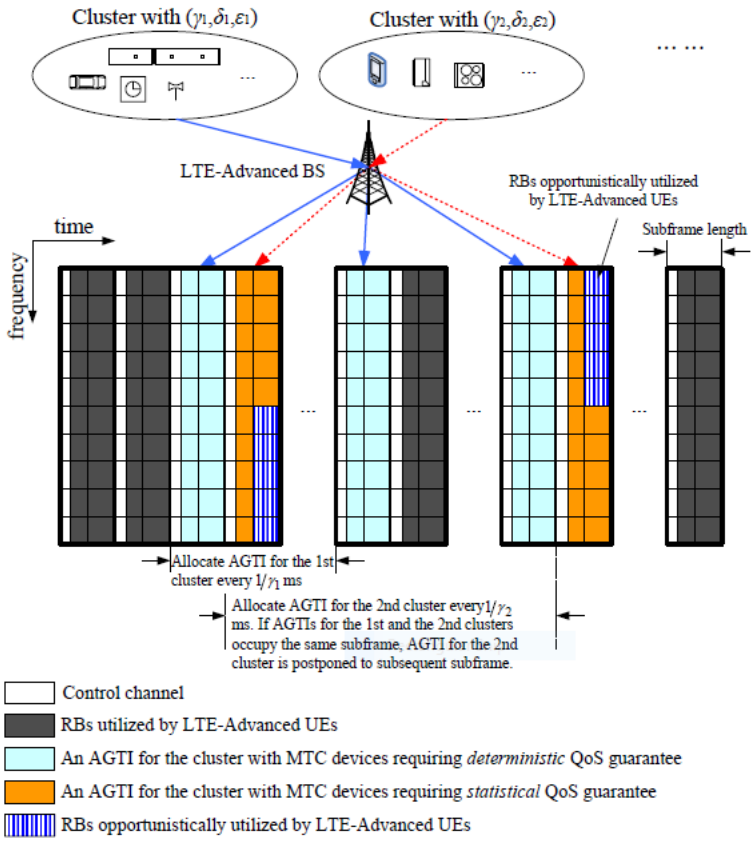


Figure 2.5: Lien *et al.*'s solution with two clusters requiring deterministic and statistical timing constraints respectively, showing how LTE-A resources are allocated within each subframe. © 2011 IEEE, from [39].

This scheme is based on these three parameters: (**a**) packet arrival rate, γ; (**b**) maximum tolerable jitter, δ; and (**c**) acceptable probability that the jitter violates δ, ϵ. Clusters with a higher packet arrival rate, and with the deterministic timing constraints, have higher priority. Clusters with lower priority are scheduled to the next subframe, if two clusters occupy the same subframe, as shown in Figure 2.5.

The significance of [39], is that multiple clusters of devices with varying timing constraints can be accommodated, while ensuring that ϵ is always satisfied, and devices with statistical timing constraints are controlled in times of high traffic volume.

However, [39] does not consider M2M/IoT devices with a non-periodic packet arrival rate (i.e. bursty traffic), as stated in MMC challenge **M3** (i.e. how to accommodate periodic and non-periodic IoT traffic), meaning that [39] is not an ideal solution for the IoT.

2.3.1.4 Clustering with Adaptive Massive Access Management

Clustering with an online estimation based algorithm for adaptive access management was proposed in [91]. The major difference over [39], is that this solution fully solves MMC challenge **M3**, in addition to MMC challenge **M2**. Utilising a wired or wireless connection between the MTC Gateway (MTCG) and the eNodeB, alongside truncated channel inversion (a power control scheme from [92]) was also proposed to simplify the radio resource allocation problem due to frequency-selective fading.

The eNodeB allocates devices to the k-th cluster based on their QoS requirements (such as in [39]), and allocates an Access Grant Time Interval (AGTI) for the k-th cluster every T_k TTIs based on their AGTI allocation periods. The eNodeB computes the upper bound queue threshold, B_H^k, which is the most access requests that can be served by the eNodeB within the maximum tolerable delay d_k. Each device within a cluster is allocated a certain number of RBs, with the limit of 1 packet in an AGTI. If the overflow probability (queue length $> B_H^k$) is higher than the 'acceptable probability that the delay violates d_k', ϵ_k, then the service rate is increased by decreasing T_k. As with the solution in [39], the cluster with a lower priority is postponed to the next subframe.

The overflow probability is estimated by using past online observations. If the average queue length increase per TTI is higher than the achievable average increase of the cluster queue, then the service rate is increased. Spectrum efficiency can be improved by reducing the service rate when the overflow probability is consistently lower than a small overflow probability threshold P_T^k.

The significance of [91] is that the proposed management scheme achieves a lower probability of QoS violation for the higher priority cluster, while it is higher for the lower priority clusters, compared to two other management schemes of Massive Access Management (MAM) in [39] and the Effective Bandwidth Based Period Scheduling (EBBPS) in [93].

However, the weighted mean delay is higher because higher priority clusters are allocated resources over the lower priority clusters, and indirect communication through a MTCG is used over direct communication with the eNodeB. Each device would also need to rely on the MTCG to reliably transmit their data to the eNodeB.

2.3.1.5 Clustering with Data Buffering

Clustering was also proposed in [68], with data buffering to minimise the control signalling overhead, and to increase the efficiency of data transmission. Each cluster would consist of devices that are close to each other (as shown in Figure 2.6), where the CHs would be the only devices that communicate directly with the eNodeB. Cluster members would communicate with each other and their CH via D2D communication, solving MMC challenge **M2**. Data from the multiple devices in each cluster is buffered at the CH, so that CHs use only a single set of control signalling. This would reduce the control signalling overhead and would also preserve the scheduling order based on priority within each cluster.

The significance of [68] is that when the amount of data to be transmitted by each device is very small, their proposed solution can accommodate tens of thousands of devices per eNodeB compared to current procedures in LTE-A. The ratio of control signalling overhead vs. data transmitted, is also much lower than other solutions used in their comparison. As the duty cycle time increases, then the number of devices that can be accommodated by an eNodeB also increases as a linear relationship.

While clustering also increases the reliability of data transmission, each member must however rely on the CH to reliably transmit their data to the eNodeB. The delay also increases when using CHs for indirect communication with the eNodeB.

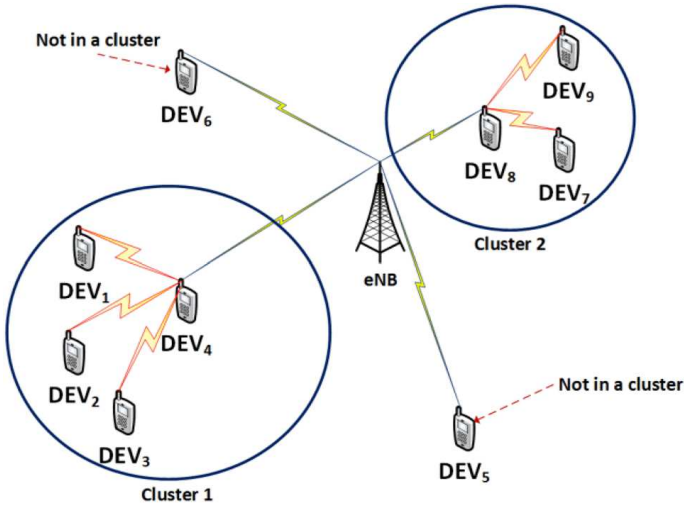


Figure 2.6: Plachy *et al.*'s solution with data buffering. Devices physically located close to each other are placed into clusters. © 2015 IEEE, from [68].

2.3.1.6 Partial Clustering for Higher Energy-Efficiency

A partial clustering solution was proposed in [94] to further deal with EE, to solve MMC challenge **M4** (i.e. devices need to be as energy efficient as possible, since replacing batteries may be impractical). Their clustering solution is fairly similar to others within literature, but the clustering is regarded as 'partial' because only devices that are far away from the eNodeB are grouped into clusters.

Azari proposes that a frame is split into two parts (as shown in Figure 2.7), one for inter-cluster communication, and the other for communication between the CH and the eNodeB. Within a cluster, the use of Carrier Sense MA-Collision Avoidance (CSMA/CA) is proposed, while TDMA is used between the CH and the eNodeB.

The inter-cluster communication part is further split into 'phases', where a cluster member is only 'active' during its assigned phase for data transmission to the CH. A portion of cluster members is allowed to compete using CSMA/CD within a phase, which reduces the collision probability and therefore saves energy. The remaining cluster members are in 'sleep mode' during this particular phase, meaning energy is not wasted by trying to sense if the channel is currently available for data transmission.

The significance of [94] is that the proposed MAC protocol leads to less collisions and less time is spent being 'idle' during communication between cluster members. The energy usage is also lower for each device if the transmission power threshold is at an optimal value, but at the cost of increased delay.

However, this solution is not ideal if EE is considered the most important parameter, as the CH will run out of energy faster than the cluster members, meaning that there would be no avenue for communication between a cluster and the eNodeB. This situation

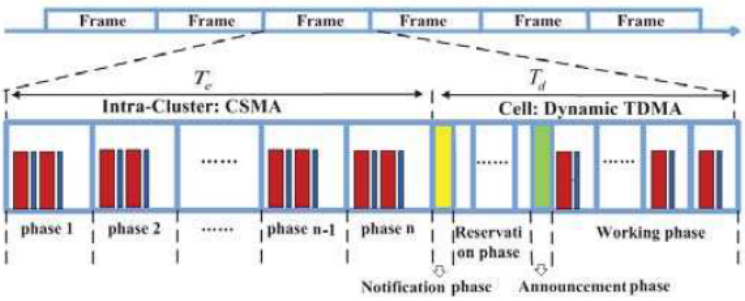


Figure 2.7: Azari *et al.*'s partial clustering solution, where a frame is split into 2 parts for inter-cluster communication and CH to/from the eNodeB. © 2014 IEEE, from [94].

is known in wireless sensor networks, as the 'hot spot problem' [5]. The network lifetime can be improved if cluster members take turns to be the CH to improve the lifetime of each cluster, and therefore the entire network.

2.3.1.7 Optimal Number of Clusters for Lower Energy Consumption

A clustering design was proposed in [45] to minimise the energy consumption of M2M devices and maximise the network lifetime [48], since M2M primarily drives the IoT [50]. The optimum number of clusters are found to maximise the network EE, where CHs are selected, and the role of CH is also rotated. The proposed design takes into account transmission and circuit energy consumption.

A device is selected as the CH of a cluster, based on a 'Cost Function', where the system chooses the best CH to minimise energy consumption. A device is selected as the CH if its Cost Function is the lowest, since it will have the lowest energy consumption for intra/inter-cluster communication. The role of CH is rotated after a certain time T, so that the network lifetime is maximised, as the increased energy consumption of CHs are shared amongst all devices in a cluster.

The significance of [45], is that the proposed clustering design achieves a higher residual energy compared to the non-clustered case, over 300 days. When the 'probability of a device to become a CH', $p = 0.7$, the residual energy is closer to the non-clustered case, as more devices within the network become CHs. All 6000 devices also run out of energy faster than all other cases, excluding the non-clustered case. When p is closer to the optimal $p = 0.1577$), then the residual energy after 300 days is close to the optimal case. Some of the 6000 devices run out of energy faster than the optimal case, but all devices run out at approximately the same time. There was also an improvement of 50% in network lifetime compared to the non-clustered case.

However, as with the previously proposed clustering solutions, utilising CHs within an OMA network will still lead to increased delay, and the devices within a cluster would need to rely on the CH to reliably transmit their data to the eNodeB.

2.3.1.8 Optimum Cluster Head Selection

A clustering design was also proposed in [46], based on an approach called Optimum Selection of CHs (OSCH), to minimise the energy consumption of M2M devices and maximise the network lifetime. The following system model (as shown in Figure 2.8) is

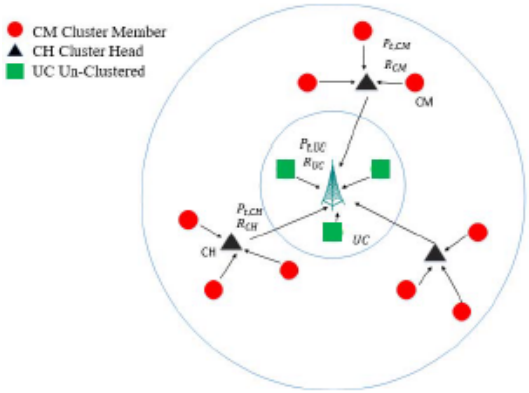


Figure 2.8: El-Feshawy *et al.*'s OSCH system model. © 2018 IEEE, from [46].

considered, where devices are arranged into clusters and the optimum CH is selected. CHs will forward collected cluster member data to the eNodeB.

Devices are only arranged into clusters if their transmission powers are lower than the threshold power value P_{th}, meaning that devices located close to the eNodeB are not clustered (similar to [94]). A genetic algorithm is applied to obtain the optimal p (as in [45]). The network is divided into 8 sectors with a designated CH in 7 of them, with the residual CHs in the 8th sector. Each sector is further divided into sub-sectors, based on the required number of CHs. The devices with the lowest transmission power in each sub-sector are designated as CHs, and all members transmit their data to their respective CHs.

The significance of [46], is that the proposed clustering design achieves a higher residual energy level after 50 days compared to the clustering design in [45] and the non-clustered case. The higher the value of p, the closer the residual energy level is to the non-clustered case. The system lifetime is maximised when p equals to the optimal p. The residual energy level increases at a higher rate for the OSCH design (≈ 1.833 times) compared to the cluster design in [45].

However, as with the previously proposed clustering solutions, utilising CHs within an OMA network will still lead to increased delay, and the devices within a cluster would need to rely on the CH to reliably transmit their data to the eNodeB.

2.3.1.9 Resource Allocation with Clustering and Q-learning

A Q-learning algorithm accompanied with the K-means Clustering algorithm (to be covered later in Section 2.4) was proposed in [43], for resource allocation of MTCDs. Q-

learning is a reinforcement learning technique, where an agent learns through trial and error, the effect its actions and decisions have on the surrounding environment [95]. The agent makes further decisions based on a reward function, R, to maximise its long-term rewards.

It is proposed that each controller (CHs in this case) select slots for their cluster members, such that the selected slot is not the same slot as chosen by a neighbouring cluster, on the same frequency. Each slot has an associated Q value for every CH, where the Q value represents the preference for a CH to select that particular slot. When there is NAK feedback, R is negative, while if there is ACK feedback, R is positive. This results in the controllers learning which slot to use for further data transmission according to the slot that has the current maximum Q value.

The significance of [43] is that the proposed Q-learning algorithm reached a much higher convergence success probability (≈ 0.6 vs ≈ 0.1), quicker in ≈ 20 slots in time, compared to slotted ALOHA and channel-based allocation (CBA) techniques. It has been shown that the negative R should be lower, to reduce the average convergence time. The average convergence time also stays linear if the positive and negative rewards are of similar value. However, this solution is not ideal for MMC/IoT, as there would not be enough time slots or frequency bands to accommodate all devices, even if MMC challenge **M2** is solved.

2.3.2 Ultra-Reliable Low-Latency Communication (URLLC)

As stated in URLLC challenge **U1** in Subsection 2.2.4 (on page 20), the current LTE-A TTI of 1 ms is not sufficient enough, even if the 10 contention-free based preambles are used during the RACH RA procedure. This is because the end-to-end (E2E) delay actually consists of many different delays, such as from the radio access network, the core network, computing processing, etc; where the radio access network delay is typically a small contributor [87].

The latency (**L1**–**L3**) can be reduced in the following ways (as detailed in Subsubsection 2.2.3.2) [87,96]: (**L1.**) increasing the subcarrier spacing, so that the durations of the 14 OFDM symbols are reduced; (**L2.**) reducing the number of OFDM symbols within a TTI. The TTI duration with 'mini-slot' lengths can be reduced by a further 7, 3.5 and 2 times for symbol lengths of 2, 4 and 7, respectively, and; (**L3.**) classifying an OFDM symbol in TDD mode as 'flexible', meaning that a device can operate in both the DL and

UL direction within the same TTI (i.e. DL at the beginning and UL at the end).

The reliability (**Re1**–**Re3**) can be increased in the following ways [87, 96]: (**Re1.**) selecting a MCS that ensures a BLock Error Rate (BLER) of $\leq 0.0001\%$ (i.e. for an ultra-reliable probability of $\geq 99.999\%$); (**Re2.**) enlarging the control resources or shortening the control information size, and; (**Re3.**) using slot aggregation or repetition, by sending the same packet over the following 1–7 TTIs.

2.4 Clustering Algorithms

Two of the most popular clustering algorithms are *Hierarchical Clustering* and *K-means Clustering*, where K-means is preferable because of the lower time complexity, faster convergence, simplicity, reliability, the ability to form new clusters after each iteration, and better suitability for larger networks (i.e. sensor or ad-hoc) [97–99]. However, K-means cluster formation can be affected by these factors: (**a**) the initial choice of the cluster centroid locations; and (**b**) the number of clusters k that must be specified prior to running the algorithm [100].

K-means Clustering has a computational complexity of $\mathcal{O}(Ikna)$, where I is the number of iterations required to converge, n is the number of data points, and a is the number of attributes (e.g. x- and y-coordinate). This complexity is essentially linear with n, as I is usually a small number. Hierarchical Clustering has a complexity of $\mathcal{O}(n^3)$, but can be reduced to a quadratic complexity of $\mathcal{O}(n^2 \log n)$, if the proximity matrix data is stored as a sorted list [97].

2.4.1 Agglomerative Hierarchical Algorithm

The steps for Agglomerative Hierarchical Clustering is given in **Algorithm 2.1** [97]. The Euclidean distance between two devices (shown in Equation (2.1)) can be used to compute the proximity matrix (shown in Equation (2.2)), where N_{ip} and N_{iq} are the locations for device p and q, respectively, and i is the dimension of the data set.

$$d_{p,q} = \sqrt{\left(\sum_{i=1}^{2} \left|N_{ip} - N_{iq}\right|^2\right)} \tag{2.1}$$

Algorithm 2.1: Hierarchical Clustering

1. Compute the proximity matrix, according to Equation (2.2).
2. Merge the closest two clusters (all devices are initially standalone clusters).
3. Update the proximity matrix with the proximity between the newly formed cluster and the original clusters.
4. Repeat steps 2-3 until there is only one cluster left.

$$\begin{bmatrix} d_{p,p} & d_{p,q} \\ d_{q,p} & d_{q,q} \end{bmatrix} = \begin{bmatrix} 0 & d_{p,q} \\ d_{p,q} & 0 \end{bmatrix} \tag{2.2}$$

The merging of the closest two clusters in Step 2 of **Algorithm 2.1**, can be done with either of the three cluster proximity methods of MIN, MAX and Group Average (as shown in Figure 2.9). **MIN** involves merging the two closest clusters which have the shortest Euclidean distance between them, while **MAX** involves merging the two closest clusters which have the farthest Euclidean distance between them. **Group Average** is the middle ground of MIN and MAX, represented as shown in Equation (2.3), where proximity(C_a, C_b) is the average proximity between cluster C_a and C_b, proximity$(\mathbf{x}, \mathbf{y})$ is the proximity between the points in cluster C_a and C_b, and m_a and m_b are the sizes of cluster C_a and C_b, respectively.

$$\text{proximity}(C_a, C_b) = \frac{\sum_{x \in C_a, y \in C_b} \text{proximity}(\mathbf{x}, \mathbf{y})}{m_a \times m_b} \tag{2.3}$$

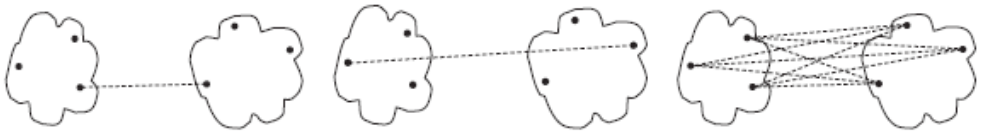


Figure 2.9: The three Hierarchical Clustering proximity methods of MIN, MAX and Group Average, respectively. © 2006. Reprinted from [97], by permission of Pearson Education, Inc., New York, USA.

2.4.2 K-means Algorithm

The steps for traditional K-means Clustering is given in **Algorithm 2.2** [2,43], where an example of the first 4 iterations of the K-means Algorithm, is shown in Figure 2.10.

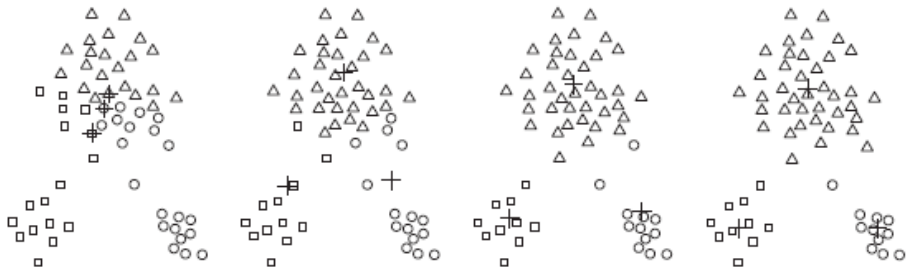
Algorithm 2.2: Traditional K-means Clustering

1. A k number of centroids are chosen at random locations within the network.
2. All devices are assigned to the closest centroid, according to Equation (2.5).
3. Each centroid location is recalculated based on the average of the device locations within that particular cluster, according to Equation (2.6).
4. Repeat steps 2 and 3 until the centroid locations do not change.

The Euclidean distance between a device and a centroid can be used for the formation of clusters, which is shown in Equation (2.4), where N_{in} is the device location and C_{im} is the centroid location; and in Equation (2.5) for two dimensions.

$$d_{n,m} = \sqrt{\left(\sum_{i=1}^{2} \left|N_{in} - C_{im}\right|^2\right)} \tag{2.4}$$

$$d_{n,m} = \sqrt{\left(\left|N_{xn} - C_{xm}\right|^2 + \left|N_{yn} - C_{ym}\right|^2\right)} \tag{2.5}$$



(a) Iteration 1. (b) Iteration 2. (c) Iteration 3. (d) Iteration 4.

Figure 2.10: K-means Clustering example, showing the first 4 iterations with the '+' indicating the centroids. © 2006. Reprinted from [97], by permission of Pearson Education, Inc., New York, USA.

The new centroid location, C_m, are then determined by Equation (2.6), where U_m is the set of devices within the m-th cluster.

$$C_m = (C_{xm}, C_{ym}) = \frac{\sum_{n \in U_m} (N_{xn}, N_{yn})}{\|U_m\|} \tag{2.6}$$

2.4.3 Density-Based Spatial Clustering of Applications with Noise (DBSCAN) Algorithm

The steps for DBSCAN are given in **Algorithm 2.3**, where **core points** are the centre point of a given neighbourhood of points (determined by a radius parameter *Eps*), where the number of points (including the core point) exceeds a certain threshold *MinPts*; **border points** are non-core points that falls within the neighbourhood of a core point, and; **noise points** are all other points. Examples of core, border and noise points are shown in Figure 2.11 [97].

Algorithm 2.3: DBSCAN

1. All points are labelled as either core, border or noise points.
2. Noise points are eliminated.
3. An edge point in inserted between all core points that are within *Eps* of each other.
4. Each group of connected core points are separated into different clusters.
5. Each border point is assigned to one of the clusters of its associated core points.

DBSCAN has a complexity of $\mathcal{O}(n \log n)$ and $\mathcal{O}(n^2)$ in the best and worst cases, respectively. Therefore, DBSCAN has a lower complexity than Agglomerative Hierarchical and K-means, but is more suited for networks with densely populated points. Therefore, DBSCAN could be more appropriate to use for MMC Networks over K-means, but it has trouble with data of widely varying densities and data with high-dimensions. The complexity of DBSCAN is especially high for high-dimensions as the proximity of the nearest neighbours must be computed, to determine core, border and noise points for each cluster [97].

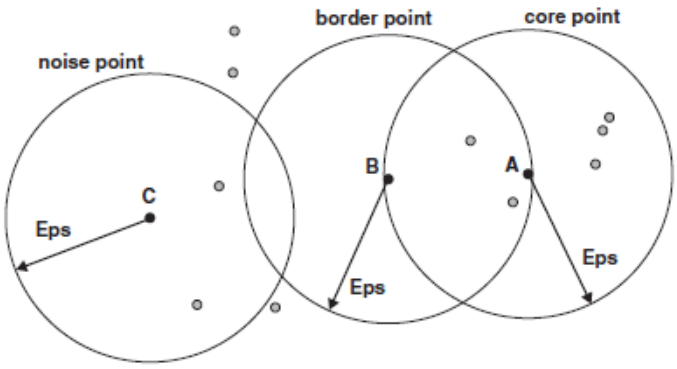


Figure 2.11: Example of core (point **A**), border (point **B**) and noise (point **C**) points in DBSCAN Clustering. © 2006. Reprinted from [97], by permission of Pearson Education, Inc., New York, USA.

2.5 Non-Orthogonal Multiple Access (NOMA)

2.5.1 Multiple Access Techniques

The throughput performances of the MA techniques of OMA and NOMA for the two scenarios of coordinated and uncoordinated allocation of resources, are compared in [69]. Coordinated resource allocation involves the eNodeB pre-allocating RBs based on its knowledge of which devices need to transmit data. The eNodeB also uses CSI to allocate RBs for the purpose of optimising the throughput. Uncoordinated resource allocation involves devices using slotted random access. RBs are not pre-allocated in this case.

NOMA commonly involves separating devices by differing power levels in the power-domain or codes in the code-domain. Superposition Coding (SC) is used at the Tx, so that devices can share spectrum. This makes NOMA more spectrally efficient than OMA, especially when channel gain differences are large. SIC at the Rx can be used to mitigate the interference from other devices and for separation of device signals [70].

The simulation results show that NOMA achieves a higher average throughput than FDMA and TDMA, if the number of devices within a network is large. This is because of the high possibility of an *overload and access problem* for OMA.

As mentioned in Subsection 1.3.2, LTE-A uses the OMA techniques of TDMA and FDMA, which has been shown to be inadequate in fulfilling the high demands of the 5G cellular network [70]. Power-domain NOMA is instead proposed as a strong candidate MA technique [51, 54, 71–73].

2.5.2 Downlink NOMA

A set of d M2M devices, can be represented as $n \in \{1, 2, 3, \ldots, d\}$. Each device can then be referred individually by D_n. The superimposed signal transmitted by the eNodeB is shown in Equation (2.7), where α_n is the power allocation coefficient for D_n, P_T is the DL power budget, and $x_n(t)$ is the individual information conveying OFDM waveform.

$$x(t) = \sum_{n=1}^{d} \sqrt{\alpha_n P_T}\, x_n(t) \tag{2.7}$$

The received signal for D_n is shown in Equation (2.8), where h_n is the Rayleigh fading channel coefficient for the link between the eNodeB and D_n, and $w_n(t)$ is the Additive Gaussian White Noise (AWGN) at D_n, with spectral noise power density N_0.

$$y_n(t) = x(t)\, h_n + w_n(t) \tag{2.8}$$

Since h_n is the channel gain coefficient, which includes path loss, fading and shadowing. The *Rayleigh fading channel* [101] can be modelled as $\mathcal{CN}\big(0, 1 \,/\, (d_n)^{\alpha}\big)$, with $\mu = 0$ and $\sigma^2 = \big(1 \,/\, (d_n)^{\alpha}\big)$. h_n can then be found by Equation (2.9), where α is the path loss exponent, d_n is the distance (in kilometres) between D_n and the eNodeB, and $X = x + iy$. X is a complex normally distributed random variable with $\mu = 0$ and $\sigma^2 = 2$. X is normalised to $\sigma^2 = 1$ and h_n is normalised to $\sigma^2 = \big(1 \,/\, (d_n)^{\alpha}\big)$, by multiplying with $\sqrt{1/2}$ and $\sqrt{1/(d_n)^{\alpha}}$, respectively.

$$h_n = X \sqrt{\left(\frac{1}{2(d_n)^{\alpha}}\right)} \tag{2.9}$$

As mentioned in Subsection 2.5.1, SIC can be used to mitigate the interference from other devices and for separation of device signals. The device signals can be decoded during the SIC process in two different orders, such as (**i**) from the strongest signal (i.e. the user with the highest allocated power) to the weakest signal and (**ii**) from the weakest signal to the strongest signal. As shown in [71–73], the most optimal SIC decoding order is the former case, since the achievable rate at the strong user while decoding the weak user's signal is greater than the achievable rate of the weak user while decoding its own signal (and treating the strong user's signal as noise).

NOMA in the DL was actually included in 3GPP Release 13 for LTE-A Pro [102], but as a special 2-user case termed *Multi User Superposition Transmission* (MUST).

2.5.3 Uplink NOMA

The superimposed signal received by the eNodeB is shown in Equation (2.10), where $x_n(t)$, h_n and $w(t)$ are defined as in the DL.

$$y(t) = \sum_{n=1}^{d} x_n(t)\, h_n + w(t) \tag{2.10}$$

2.5.3.1 Multi-User Shared Access (MUSA) for IoT

A grant-free code-domain NOMA scheme was proposed in [50], called *Multi-User Shared Access*, to support the IoT. It involves spreading the data of users with a family of complex spreading sequences of a short length. The signals are then superimposed at the Rx, so that SIC can be carried out to mitigate the interference from other users. This spreading sequence is designed to cope with the overloading of users and to simplify the SIC process at the Rx. The spreading codes can be autonomously chosen by users, thereby moving resource coordination to the Rx side rather than at the eNodeB.

The overloading ratio (**OR**) is shown in Equation (2.11), where K is the number of users to be overloaded and L is the length of the complex spreading code applied to each modulated symbol, to be transmitted over L resources.

$$\mathbf{OR} = \frac{K}{L} \tag{2.11}$$

The number of available codes is $(M^2)^L$, where M is the number of available values of each element of a complex M-ary spreading code (as shown in Figure 2.12, for $M = 2$ and $M = 3$, respectively).

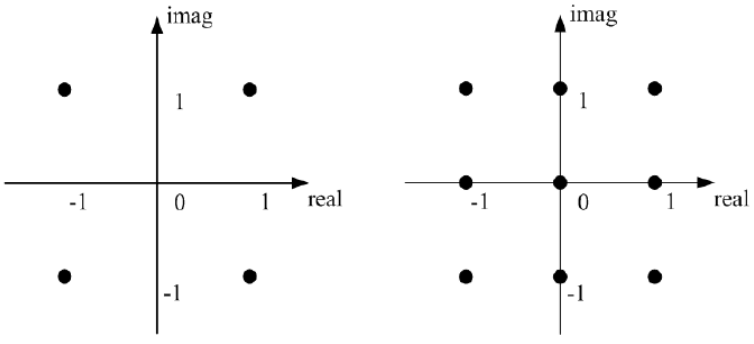


Figure 2.12: Elements of a complex M-ary spreading code for $M = 2$ and $M = 3$, respectively. © 2016 IEEE, from [50].

The significance of [50] is that MUSA can facilitate high overloading ratios of users (i.e. "the fundamental principle of NOMA is to serve multiple users at the same channel use" [55]). Having a higher L enables support of higher overloading ratios with a lower BLock Error Rate (BLER). A BLER of $< 1\%$ can be easily achieved when $M = \{2, 3\}$ and $L > 8$, compared to the binary pseudo-random noise (PN) sequence, which cannot even with a $L = 4$. MUSA with $M = 3$ was also compared to a random Guassian complex spreading code (the expected theoretical optimum), with slightly worse performance. MUSA was also compared to LTE-A in system level simulations, and achieved a lower BLER for most SNR points even with overload ratios of 200, 300 and 400%. MUSA also had a lower packet loss rate at higher traffic loads than OFDMA.

2.5.4 User Fairness

The fairness index — representing how symmetrical user's individual data rates are — is given in Equation (2.12), with $k = \{1, 2, \ldots, K\}$ number of users, and R_k is the data rate of the k-th user. A **Fairness** of 1 indicates maximum symmetrical data rates, while a **Fairness** of 0, indicates maximum asymmetrical data rates [70, 103, 104].

$$\textbf{Fairness} = \frac{\left(\sum_{k=1}^{K} R_k\right)^2}{K \sum_{k=1}^{K} R_k^2} \in [0, 1] \tag{2.12}$$

NOMA has been shown to have a better trade-off between system throughput and user fairness compared to OMA [52, 55, 71, 104–106], because strong channel gain users can utilise the same resources as weak channel gain users, which is otherwise inaccessible in OMA [72]. OMA also tends to prioritise strong channel gain users rather than weak channel gain users (usually located at the cell-edge), therefore the fairness index would be low in OMA.

NOMA and OMA can achieve a **Fairness** of 1, if the objective is to maximise the system throughput and the number of users is small. When the number of users is large, OMA can achieve a **Fairness** of 1, by having low data rates or using a fair scheduling policy. However, the feedback and control signalling overhead will be significant in the low data rate region, leading to excessive delays in LTE-A. Prioritising users requiring low data rates will also lead to the starvation of resources for 5G eMBB devices, since M2M/IoT devices will be numerous compared to eMBB in 5G.

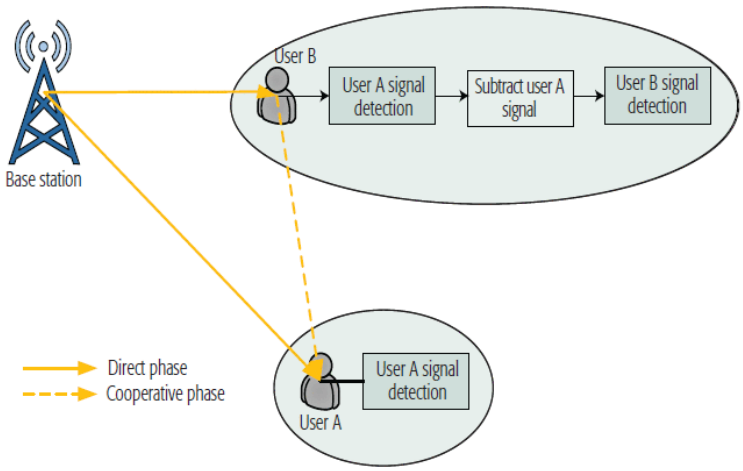


Figure 2.13: 2-device cooperative NOMA with the 2 phases of the usual NOMA and then the cooperative phase © 2017 IEEE, from [55].

2.5.5 Cooperative NOMA

A cooperative NOMA scheme was proposed in [53], where the prior information that strong channel users have of the other users within the network, is exploited to increase the weak users' reception ability. For example, in a 2-user NOMA system, the strong user (e.g. **device A**) can act as a relay for the weak user (e.g. **device B**). **Device A** can send **Device B**'s portion of the superimposed message it has decoded during the SIC process, to **device B**. As a result, there are now two copies of **device B**'s message, and therefore this increases **device B**'s own reception ability.

The cooperative NOMA scheme consists of two phases, where the first phase is the usual NOMA, while the second phase is the 'cooperative phase', which is shown in Figure 2.13. It is proposed that short range communications such as Bluetooth, ultra-wideband (UWB), D2D, etc, should be used for the cooperative phase, as extra time slots would be needed otherwise. For n-users within a network, the d-th user receives copies of its message from all the users that have stronger channel gains.

The significance of [53] is that cooperative NOMA achieves a lower outage probability than non-cooperative NOMA and OMA, which confirms that the weak user in the network has a better reception ability than the comparable schemes at various SNR points. It was also shown that cooperative NOMA achieves a much higher bits per channel use (BPCU) than the other schemes, when $R_1 = R_2$. However, it is worth noting that cooperative NOMA has increased system complexity over regular NOMA, which is why

a more practical alternative called 'User Pairing' NOMA, was suggested.

This proposed cooperative NOMA scheme can be improved (in terms of spectral efficiency) if full-duplex mode (i.e. 2 antennas for simultaneous UL/DL) is used instead [107]. However, full-duplex mode suffers from 'self-interference', where the transmitted signal interferes with the received signal on the same transceiver [108–112].

2.5.6 User Pairing NOMA

The impacts of 'User Pairing' when utilising NOMA, was investigated in [54]. A hybrid MA system was proposed, where users are sorted into groups. NOMA is utilised within each group, while OMA is utilised between groups and the eNodeB.

Consider an example 2-user DL NOMA cluster with devices D_1 and D_2, where $|h_2|^2 < |h_1|^2$. OMA would allocate power as $P_1 = P_2$, while NOMA would allocate power as $P_1 < P_2$. Assuming that $|h_2|^2 = 10$ dB, $|h_1|^2 = 40$ dB, $P_1 = 0.3$ W, and $P_2 = 0.4962$ W, NOMA is shown to achieve $\approx 61.76\%$ higher sum throughput than OMA. If $|h_2|^2 = 30$ dB instead, then NOMA achieves $\approx 16.06\%$ higher sum throughput than OMA, which is worse than the case where $|h_2|^2 = 10$ dB and $|h_1|^2 = 40$ dB.

Typically, the data clustering algorithms featured in Section 2.4 would cluster users in terms of their similarity to other users (i.e. they would have similar channel gains). This would result in a reduced NOMA sum throughput advantage over OMA, as elaborated in the 2-user DL NOMA cluster example as above. It is therefore more beneficial to pair users that have a big difference in their Rayleigh fading channel gains as the strong user can utilise the bandwidth allocated to the weak user, which is otherwise inaccessible when using OMA [71]. A higher sum throughput naturally leads to a higher spectral efficiency.

2.5.7 Clustering and Power Allocation

A common goal for a network is to maximise the network throughput while under various constraints. One such constraint is allocating P_n in an optimal manner, while $\sum_{n=1}^{d} P_n \leq P_T$. This would require an exhaustive search of the many different combinations of power allocations, which would be very impractical. Therefore, a suboptimal user pairing and an optimal power allocation scheme, was proposed in [76]. Equations were formulated based on an optimisation problem with transmit power constraints, minimum throughput constraints and SIC constraints. Each set of power allocation equations

has a corresponding set of conditions that must be satisfied.

For the SIC process to be successful, there are necessary power constraints as shown in Equation (2.13), where P_{tol} is the minimum power difference required between the decoded signal and the rest of the signals within a NOMA cluster of size-m, to be decoded from $y_n(t)$.

$$P_n \left|h_{n-1}\right|^2 - \sum_{i=1}^{n-1} P_i \left|h_{n-1}\right|^2 \geq P_{tol},\ n = 2, 3, \ldots, m \tag{2.13}$$

The strongest channel gain user within a m-user NOMA cluster has a power allocated as shown in Equation (2.14), where $\delta = P_{\text{tol}} / \left|h_1\right|^2$ and P_t is the DL NOMA cluster power budget. $\delta \to 0$ when $\left|h_1\right|^2$ is high, so Equation (2.14) can be approximated as shown.

$$P_1 \leq \frac{P_t - \delta}{2^{m-1}} - \frac{\delta}{2^{m-2}} - \cdots - \left(-\frac{\delta}{2}\right) \approx \frac{P_t}{2^{m-1}} \tag{2.14}$$

2.5.7.1 Downlink

The proposed suboptimal user pairing scheme is shown in Figure 2.14, where the users are arranged from strongest user (UE1) to weakest user (UE12) according to the NOMA principle. The users within the network are arranged into κ clusters based on the number of α strong users. If $\alpha < d/2$, then $\kappa = \alpha$, while if $\alpha \geq d/2$, then $\kappa = d/2$. Users within the network are allocated into Cluster $1, 2, ..., \kappa$, as given in Algorithm 1 on page 6331 of [76].

The accompanying power allocation sets of equations were derived from checking 2^{m-1} Lagrange multiplier combinations that satisfy the Karush-Kuhn-Tucker (KKT) conditions of the optimisation problem. Normally 2^{2m} Lagrange multiplier combinations need to be checked, which has a high complexity. However, since $P_n > 0$, then not all combinations need to be checked. [76] presents a 2^{m-1} sets of equations with accompanying necessary conditions in Table 1 on page 6334 of [76]. These necessary conditions are generally in the form as shown in Equation (2.15) and (2.16).

$$P_n \gamma_n - (\varphi_1 - 1)\left(\sum_{j=1}^{n-1} P_j \gamma_n + \omega\right) > 0,\ n = 1, 2, \ldots, m \tag{2.15}$$

$$\left(P_n - \sum_{j=1}^{n-1} P_j\right)\gamma_{n-1} - P_{tol} > 0,\ n = 2, 3, \ldots, m \tag{2.16}$$

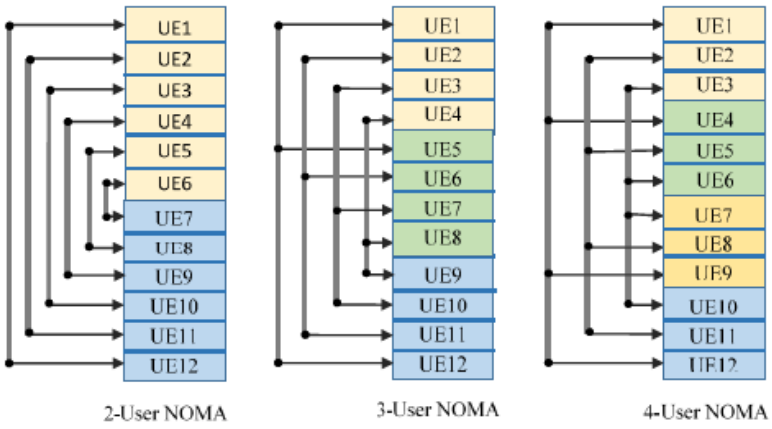


Figure 2.14: Ali *et al.*'s user clustering solution for 2/3/4-user NOMA with $n = 12$ active DL users in the network. © 2016 IEEE, from [76].

The results in [76] show that: **(i)** if the $|h_n|^2$ of all users are similar to each other, then it is preferable to have larger cluster sizes for high $|h_n|^2$ and smaller cluster sizes for low $|h_n|^2$; **(ii)** if $\alpha = \kappa$, then NOMA achieves maximum throughput gain over OMA; **(iii)** cluster size is not that significant if $\alpha \geq 50\%$ of the number of users within the network, and; **(iv)** η_1 is always maximised when using NOMA, but the proportionality of throughputs within the network can be improved by increasing R_i. However, the maximum allowed R_i is limited to the $|h_n|^2$ of the weak user within a cluster.

2.5.7.2 Uplink

The proposed suboptimal user pairing scheme is shown in Figure 2.15, where the users are arranged from strongest user to weakest user, and into κ clusters based on the number of α strong users, as in the DL. Users within the network are allocated into Cluster $1, 2, ..., \kappa$, as given in Algorithm 1 on page 6331 of [76].

The accompanying power allocation sets of equations were derived from checking 2^m Lagrange multiplier combinations that satisfy the KKT conditions of the optimisation problem. Normally 2^{3m-1} Lagrange multiplier combinations need to be checked, which has a high complexity. However, since $P_n > 0$, then not all combinations need to be checked. [76] presents 3 sets of equations for each m with accompanying necessary conditions in Table 2 on page 6335 of [76]. These necessary conditions are generally in the form as shown in Equation (2.17)–(2.19).

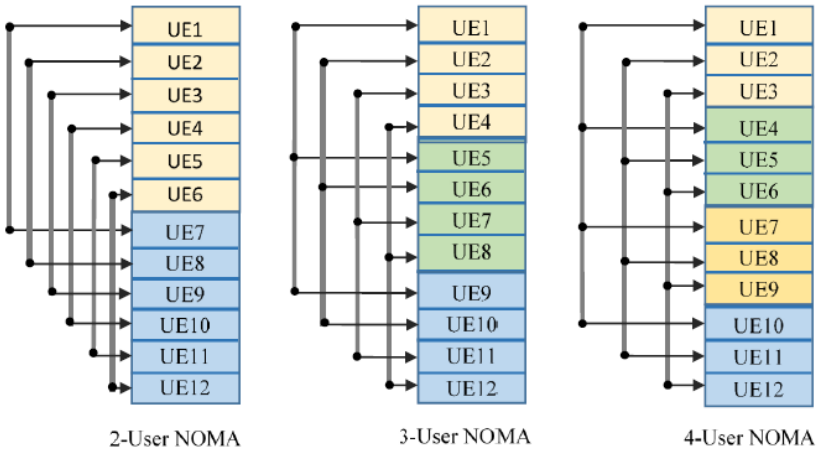


Figure 2.15: Ali *et al.*'s user clustering solution for 2/3/4-user NOMA with $n = 12$ active UL users in the network. © 2016 IEEE, from [76]

$$(\mathbf{C_1^m}) \rightarrow P_n\gamma_n - \sum_{j=n+1}^{m} \phi_n P_j \gamma_j - \phi_n \omega > 0,\ n = 1, 2, \ldots, m \tag{2.17a}$$

$$(\mathbf{C_2^m}) \rightarrow P_n\gamma_n - \sum_{j=n+1}^{m} P_j \gamma_j - P_{\text{tol}} > 0,\ n = 1, 2, \ldots, m-1 \tag{2.17b}$$

$$(\mathbf{C_1^m})\ n = 1, 2, \ldots, m \wedge \neg(n = m-1),\ \textbf{AND}\ (C_2^m),\ \textbf{AND}\ P_m < P_t' \tag{2.18}$$

$$(\mathbf{C_1^m}),\ \textbf{AND}\ (C_1^m)\ n = 1, 2, \ldots, m-2,\ \textbf{AND}\ P_m < P_t' \tag{2.19}$$

As mentioned in Subsection 2.5.2, the most optimal SIC decoding order is from the strongest signal to the weakest signal. Contrary to the DL, all users but the weakest channel gain user (which may have power control applied to maintain power level distinctness) transmit at the same power level in the UL. Therefore, the strongest channel gain user is most likely to have the strongest signal, meaning that the eNodeB decodes this signal first during the SIC process [76].

2.6 Simultaneous Wireless Information and Power Transfer (SWIPT)

SIC is an essential part of NOMA, but it has a complexity of $\mathcal{O}(L^3)$, where L is the number of signals that have been superimposed [49]. This complexity is very high within an MMC/IoT network, leading to significant power consumption. Utilising 'user pairing' NOMA [54,77] (as mentioned in Section 1.3), is similar to placing the devices into clusters, but it also maximises the NOMA network sum throughput gain over OMA. User Pairing and clustering significantly reduces the SIC complexity, since L is now divided amongst the clusters. Therefore, NOMA requires clustering in large networks such as MMC/IoT.

SWIPT involves energy harvesting (EH) from the received RF signals, which effectively offsets the power consumption of SIC and thereby also increases the EE. It is currently not feasible to simultaneously harvest energy and transmit information, as it would require separated EH and Tx/Rx modules [56]. Nevertheless, it is practical if these modules are co-located on the device, where the two main SWIPT strategies (shown in Figure 2.16 [56]) are: (**a**) Power-Splitting (PS), where the received RF signal is split for EH and information decoding; and (**b**) Time-Switching (TS), where the transmission time slot is divided for EH and data transmission [57–61].

2.6.1 Downlink

The effect on the throughput of D_n, when the PS SWIPT technique is applied for EH purposes, is shown in Equation (2.20), where $1 \leq P_s \leq 1$ is the portion of the received

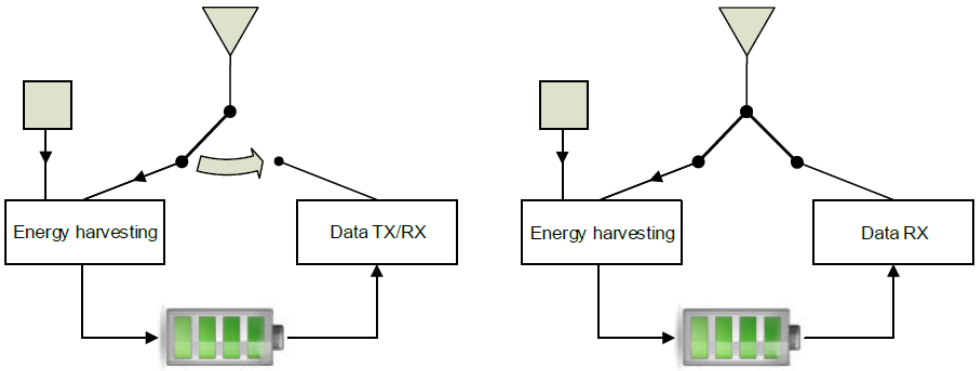


Figure 2.16: Left: Time-switching SWIPT, **Right**: Power-splitting SWIPT. © 2015 IEEE, from [56].

normalised signal used for EH. The throughput of the device decreases as P_s increases to power the DL SIC process at the device, as shown later in Figure 3.11.

$$\eta_n = B\log_2\left(1 + \frac{P_n(1-P_s)\left|h_n\right|^2}{N_0 B + \sum_{i=1}^{n-1} P_i(1-P_s)\left|h_n\right|^2}\right) \tag{2.20}$$

2.6.1.1 Cooperative NOMA with SWIPT

As mentioned in Subsection 2.5.5, cooperative NOMA has increased system complexity over regular NOMA, and therefore increased power consumption. 'User Pairing' NOMA has been suggested as a more practical alternative to reduce system complexity; however, SWIPT can also be used to mitigate the increased power consumption at the stronger device in cooperative NOMA.

PS SWIPT was used in [57], for EH during the first phase in 2-device cooperative NOMA (as shown in Figure 2.17). The stronger **B**-device uses the harvested energy from the first phase to act as a relay for the weaker **A**-device, to increase the reception ability of the weaker device. There are two common cooperative relay techniques of: (**i**) amplify-and-forward (AF), and (**ii**) decode-and-forward (DF), where cooperative NOMA naturally uses DF since **B**-devices must decode the signals of **A**-devices during the SIC process.

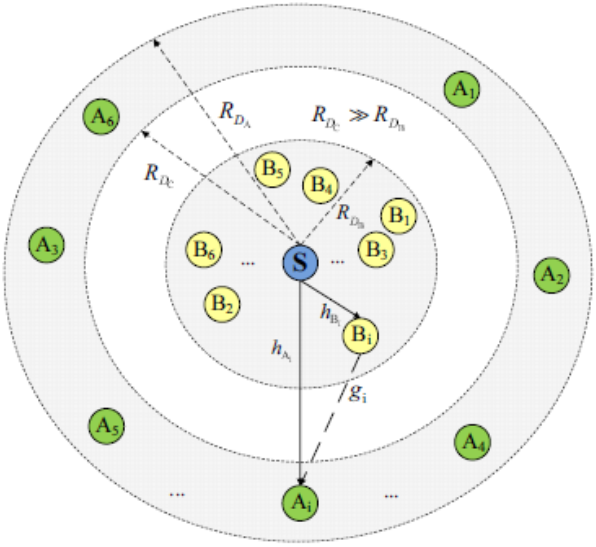


Figure 2.17: DL Cooperative NOMA scheme, where normal UL NOMA is performed between the eNodeB and strong/weak devices (**B**/**A**), but SWIPT is applied to enable the **B**-device to DF relay the **A**-device signal (from the SIC process) towards the **A**-device. © 2016 IEEE, from [57].

The **A**-device utilises Maximal-Ratio Combining (MRC) to combine the received signals from the eNodeB during the first phase and from the **B**-device during the cooperative phase. Three user pairing schemes are proposed based on the device locations relative to the eNodeB:

- random near device and random far device (**RNRF**): this does not require instantaneous CSI.
- nearest near device and nearest far device (**NNNF**): the nearest near device harvests more energy and both devices will have minimised outage probability.
- nearest near device and farthest far device (**NNFF**): the same as NNNF for the near device, but the throughput gain is prioritised by maximising the difference in channel gains between the near and far device.

The significance of [57] is that the NNNF and NNFF user pairing schemes achieve a lower outage probability than RNRF, because of the lower path loss of selecting the nearest near device rather than a random near device. For the far devices, NNNF achieves the lowest outage probability as described in the description for NNNF in the list above. NNFF achieves higher outage probability than NNNF but lower than RNRF. NNNF achieves the lowest outage probability for both non-cooperative and cooperative NOMA because of its lowest path loss. However, NNFF has a higher outage probability than RNRF for non-cooperative NOMA. The outage probability decreases more rapidly when the SNR increases for cooperative NOMA, because of higher diversity gain and the guaranteed reliable reception in the high SINR region.

2.6.1.2 Considering Rx circuitry power consumption

In [57], it was assumed that the energy required for receiving and processing data at the Rx is negligible compared to the energy required for data transmission. However, this is not the case for severely power limited applications such as wireless sensors [58].

An EH scheme is proposed in [58], combining PS SWIPT and TS SWIPT. TS SWIPT is applied first, where a slot is divided into two sub-slots. The first sub-slot features the strong user (UE1) as EH only and the weak user's (UE2) signal is carried in the RF signal from the eNodeB. In the second sub-slot, PS SWIPT is applied at UE1 (as shown in Figure 2.18), where a portion of the received RF signal, ρ, is used for energy harvesting,

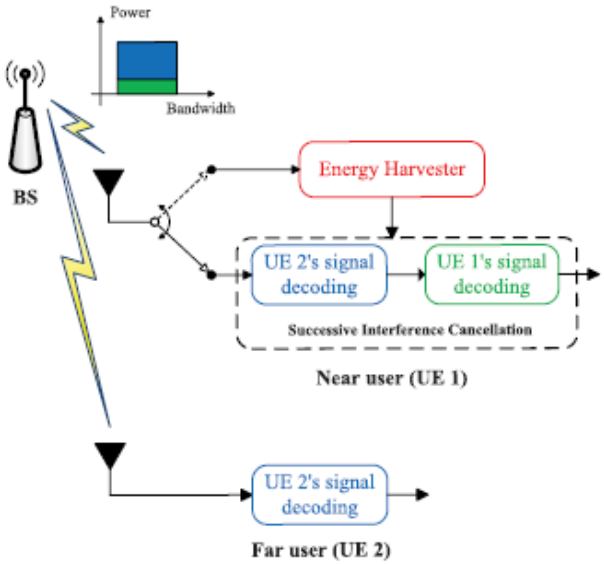


Figure 2.18: DL NOMA with an Energy Harvester at the strong user (UE1). © 2017 IEEE, from [58].

and the rest for information decoding. The total harvested energy is $\geq$ to the required energy for information decoding, is given in Equation (2.21), where $0 \leq t \leq 1$ is the duration of the first sub-slot, $(1-t)$ is the duration of the second sub-slot, and $P_x^{(y)}$ is the transmit power for the x-th UE in the y-th sub-slot. P_{SIC}, ζ and $|h_n|^2$, are defined later in Section 3.5.

$$t\zeta \left|h_1\right|^2 P_2^{(1)} + (1-t)\rho\zeta \left|h_1\right|^2 \left(P_1^{(2)} + P_2^{(2)}\right) \geq (1-t)P_{\text{SIC}} \tag{2.21}$$

The results in [58] show that for the constant decoding power consumption case of $P_{\text{SIC}} =$ 80 mW: (**i**) there is a certain cut-off point before the achievable rate for UE1, R_1, becomes 0 for TS SWIPT; (**ii**) when UE2 is closer to UE1 and the eNodeB, then R_1 and the achievable rate for UE2, R_2, are linear with each other, but there is a certain cut-off point before the achievable rate for UE1 becomes 0 for TS SWIPT, and; (**iii**) if UE1 is located too far away from the eNodeB, then PS SWIPT is infeasible.

For the dynamic decoding power consumption case of $P_r = 30$ mW and 50mW for SIC, for a total of $P_{\text{SIC}} = 80$ mW: (**i**) the results are similar with the constant decoding power consumption case for the first point in the previous list, and (**ii**) PS SWIPT is now feasible with dynamic decoding power consumption, even if UE1 is located too far away from the eNodeB.

For the dynamic decoding power consumption case of $P_r = \{10, 30, 50\}$ mW and static decoding power consumption case of $P_{\text{SIC}} = \{60, 80, 100\}$ mW: (**i**) for the combined EH scheme, all cases are feasible except for $P_{\text{SIC}} = 100$ mW. The achievable rate region becomes slightly worse as P_r and P_{SIC} increases; (**ii**) for TS SWIPT, there are various cut-off points before R_1 becomes 0, which is worse as P_r and P_{SIC} increases; (**iii**) for PS SWIPT, all cases are feasible except for $P_{\text{SIC}} = 100$ mW, and; (**iv**) if $\zeta = 0.3$, then the combined EH scheme has a cut-off point before R_1 becomes 0, but only for the static decoding power consumption case.

2.6.2 Uplink

The effect on the throughput of device D_n, when the TS SWIPT technique is applied for EH purposes, is shown in Equation (2.22), where $1 \leq T_s \leq 1$ is the portion of the normalised transmission time used for EH. The throughput of the device decreases linearly as T_s increases to power UL transmission at the device, as shown later in Figure 4.13a.

$$\eta_n = \left(1 - T_s\right) B \log_2 \left(1 + \frac{P'_n \left|h_n\right|^2}{N_0 B + \sum_{i=n+1}^{d} P'_i \left|h_i\right|^2}\right) \tag{2.22}$$

2.6.2.1 Without clustering

A greedy algorithm based on TS SWIPT was proposed in [59], using the system model shown in Figure 2.19, with a single eNodeB and an N number of EH users. Two different decoding strategies of: (**i**) fixed decoding order and (**ii**) TS, are investigated for

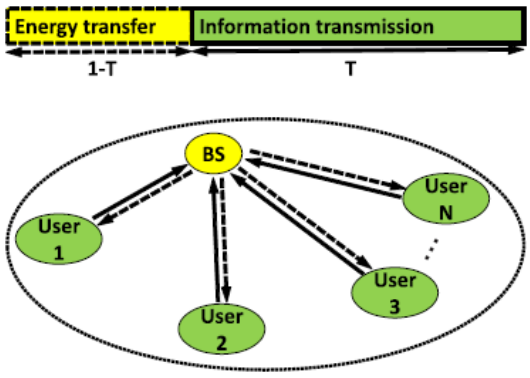


Figure 2.19: SWIPT with a single eNodeB and N EH users, where the users harvest energy in the DL to facilitate UL NOMA. © 2016 IEEE, from [59].

possibilities of increasing user fairness. The two objectives for optimising the QoS are: (**a**) Maximise achievable system throughput with minimum rate requirement; and (**b**) maximise the equal individual data rates.

Therefore, Diamantoulakis *et al.* proposes four different schemes of: low complexity **A** $\rightarrow$ (**a**) and (**i**); medium complexity **B** $\rightarrow$ (**a**) and (**ii**); medium complexity **C** $\rightarrow$ (**b**) and (**i**), and; high complexity **D** $\rightarrow$ (**b**) and (**ii**).

The **EE** of transmit power from the eNodeB, when users require an equal UL transmit rate is given in Equation (2.23), where P_0 is the eNodeB transmit power, $0 \leq T \leq 1$ is the portion of the transmission time slot used for UL transmission, and R_{eq} is the common user rate — where the rate of the weakest user is maximised.

$$\mathbf{EE} = \frac{NR_{eq}}{P_0(1-T)} \tag{2.23}$$

It is assumed that UL transmission can only use the harvested energy during $(1-T)$. Therefore, there is a trade-off between EE and NR_{eq}, where increasing the spectral efficiency (**SE**) further past the optimal EE point, leads to a decrease in EE. The harvested energy can be 'green', thereby leading to a reduction in the *carbon footprint* — the total carbon dioxide (CO_2) emissions as a result of generating this energy.

The results in [59] show that the most optimal T for a 2-user pair with similar channel gains is 0.7958, in terms of maximising the system throughput and individual throughputs. For a 2-user strong/weak user pair, the system throughput is maximised when $T = 0.8895$, however this does not maximise the individual throughputs. The individual throughputs are maximised when $T = 0.54$, but this reduces the system throughput as more power is allocated to the weak device. Using TS achieves a higher system throughput and a higher weak user throughput, compared to fixed decoding order and OMA. TS also has a higher average user rate regardless of the distance from the eNodeB. The required $(1-T)$ to maximise the system throughput, decreases when there are more users. The EE for the TS schemes are superior to OMA.

The proposed scheme in [59] achieves the goals presented at the beginning of their paper. However, their scheme would not be optimal in an IoT/MMC setting, as the increased number of non-clustered users would lead to a severely high SIC process complexity, and would lead to high SIC error propagation if the earlier devices are erroneously decoded.

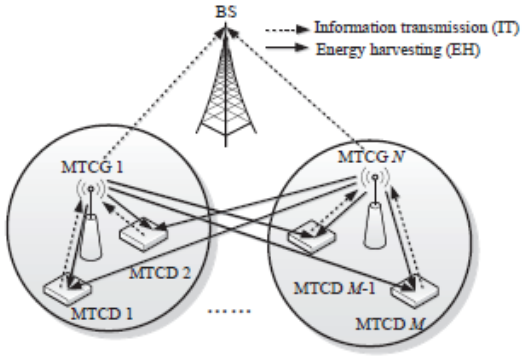


Figure 2.20: UL SWIPT while using MTCG as a relay/CH. MTCDs harvest energy from the MTCG of their respective cluster. © 2018 IEEE, from [60].

2.6.2.2 Using MTCG as relays

An energy efficient resource allocation for the IoT was proposed in [60], with the use of MTCG as a relay/CH between the eNodeB and MTCDs, with the system model shown in Figure 2.20. The proposed scheme uses TS SWIPT, where the first phase consists of MTCDs transmitting to their respective MTCG, and the second phase consisting of the MTCGs transmitting a superimposed signal to the eNodeB while the MTCDs harvest energy from their respective MTCG. This is based on the assumption that MTCGs have an abundance of energy available, and can therefore carry out more complex tasks.

The results in [60] show that the energy consumed by an MTCD decreases sharply at the beginning of the allocated time slot, and then slowly decreases when $P^C = 0$ mW (where P^C is the circuit power consumption). When $P^C = 5$ mW and $P^C = 10$ mW, the same trend occurs for the beginning of the allocated time slot, and then the energy consumed sharply increases. At the low circuit power consumption (i.e. $P^C = 4$ mW) region, the circuit power consumption is lower for NOMA compared to OMA. When $P^C = 0.5$ mW, the total energy consumed does not vary much even if the maximal transmission power of each MTCG increases from 0.1–1.1 W. When $P^C = 5$ mW, NOMA consumes more total energy than OMA if the maximal transmission power of each MTCG is ≤ 0.6 W. NOMA consumes more total energy in general compared to OMA, when the maximal transmission power of each MTCD increases from 0.1–5.1 mW. This is also true when the required payload for each MTCD increases from 9–20 kb.

2.7 Hybrid Automatic Repeat ReQuest (HARQ)

Reliability can be defined as "the probability that a certain amount of data is successfully transferred error-free from end to end, within a given deadline" [113]. As mentioned in Section 1.1, URC is one of the three major operating modes of 5G, where applications can transfer data in an ultra-reliable manner for the majority of the time [10, 12, 21]. For the URC operating mode, the ultra-reliable probability is $P_{\text{UR}} \geq 99.999\%$ [10, 21, 23, 32].

HARQ has been proven to improve the reliability of data transmission in cellular networks. There are two main types of HARQ, where HARQ can be combined with packet-combining methods, known as Chase Combining (CC) and Incremental Redundancy (IR) for Type-I and Type-II HARQ [65]. CC HARQ and IR HARQ in relation to NOMA in 5G, are discussed further in Subsection 5.2.3.

In 5G, slot aggregation (grant-based)/repetition (grant-free) can be used for HARQ purposes, to increase the reliability of data transmission, while also reducing the delay incurred from the current LTE-A HARQ process round trip time of 8 ms. Slot aggregation involves 1–7 transport blocks (TB) consisting of the same packet, being sent consecutively over 1–7 TTIs. Grant-free HARQ would significantly reduce the latency incurred if there is an erroneous packet, but is generally less efficient (in terms of throughput) than grant-based HARQ [90]. Nevertheless, the use of longer subcarrier spacing and mini-slots in 5G (TTI durations are shown in **Table 2.2**), are the key ingredients to jointly solving the challenges for URLLC (covered in Subsection 2.2.4 on page 20).

Table 2.2: The new TTI durations (in ms) in 5G

Subcarrier Spacing (kHz)	**OFDM symbols per TTI**			
	14	7	4	2
15	1	0.5	0.286	0.143
30	0.5	0.25	0.143	0.071
60	0.25	0.125	0.071	0.036
120	0.125	0.063	0.036	0.018
240	0.063	0.031	0.018	0.009

2.8 Conclusion

In this chapter, a literature review was conducted on the related work of MTC in cellular networks (LTE-A, 5G NR, frequency and time resources, LPWA cellular technologies to support the IoT, and IoT challenges), the IoT (including MMC and URLLC), clustering algorithms, and NOMA.

LTE-A is the potential candidate to enable the IoT, since cellular networks are widely deployed worldwide, offer high security, interference-free communication, QoS guarantees and are also scalable depending on the IoT use case. However, the RACH RA procedure has high potential to be a bottleneck, which eventually leads to the *overload and access problem*, and also contributes to high inefficiency from the significant feedback and control signalling overhead.

Within the literature, collision resolution, clustering and resource allocation have been proposed to deal with the *overload and access problem*. Collision resolution reduces the effect of the *overload and access problem*, but causes high delay due to collided devices being directed to compete for resources in a future RAO. It is also inefficient to reserve resources for collided devices in a future RAO, if no devices have collided during the RACH RA procedure. Clustering can reduce the feedback and control signalling overhead, and the effects of the overload and access problem. However, by itself, it is not an effective solution, as clustering also introduces delay due to the communication required between the CH and the eNodeB, and devices must rely on the CHs to reliably transmit their data to the eNodeB.

Therefore, 5G is considered the key enabler for the IoT considering that it will natively support the MMC and URC operating modes, which make up the IoT. NOMA is widely considered an enabling MA technique for 5G, and crucially avoids the *overload and access problem*. However, the essential NOMA SIC process has high complexity (and therefore high power consumption) in un-clustered networks. Therefore, clustering and NOMA are effective solutions for the IoT. To further improve the NOMA network sum throughput gain over OMA, user pairing NOMA is suggested.

Chapter 3

Downlink NOMA with SWIPT

3.1 Introduction

3.1.1 Motivations

As mentioned in Section 1.3 (on page 7), this thesis will focus on enabling the MMC operating mode, which has differing QoS requirements of massive connectivity, small packets, low data rate, high EE, and low energy consumption. One major challenge for MMC, is the possibility of the massive number of M2M devices simultaneously attempting to access network resources in an orthogonal manner. This scenario would likely lead to an *overload and access problem*, and high transmission inefficiency from the feedback and control signalling overhead. One effective solution to this challenge is by using a clustering algorithm accompanied by the NOMA technique, since clustering can reduce the feedback and control signalling overhead, and NOMA can avoid the *overload and access problem*. The two most popular iterative unsupervised learning clustering algorithms of K-means and Hierarchical are considered, where it is shown later in Subsection 3.7.3, that K-means is the preferred option.

The traditional K-means clustering algorithm is given earlier in **Algorithm 2.2** (on page 33), but the algorithm result does not benefit from the NOMA 'User Pairing' concept (as detailed in Subsection 2.5.6). Step 2 of **Algorithm 2.2**, "all devices are assigned to the closest centroid...", will cause many of the strongest channel gain devices to be assigned into the same cluster, thereby reducing the NOMA network sum throughput percentage gain over OMA. While NOMA has many benefits over OMA, the essential

SIC process however has high complexity in an MMC Network. Clustering significantly reduces this complexity and therefore the power consumption of SIC, while the PS SWIPT EH technique (as detailed in Section 2.6) can completely mitigate this power consumption. Thus, NOMA requires clustering in large networks such as MMC/IoT.

3.1.2 Contributions

Thus, an enhanced K-means algorithm with DL NOMA and SWIPT is proposed in this chapter, to fulfil the requirements of MMC devices (such as sensor nodes) and address some of the challenges mentioned above, which are used as motivations [2]. The proposed algorithm involves two subproblems of cluster determination and power allocation:

1. finding the greatest average 'silhouette value' (later defined in Section 3.4) to determine the most optimal k number of clusters, before applying K-means to the network based on the devices' locations relative to the eNodeB.

2. excluding the k strongest channel gain devices from the network to be assigned as CHs in a later step.

3. applying K-means to the remaining network to find the best solution to the cluster formation problem.

4. assigning each k excluded device to the appropriate cluster, according to the 'User Pairing' concept.

5. optimally allocating power to each device in a NOMA cluster, based on set of equations and corresponding conditions, according to the NOMA principle.

This chapter is organised as follows: Section 3.2 establishes the system model and states the assumptions. Section 3.3 details an improved K-means algorithm, while the proposed enhanced K-means with NOMA scheme is presented in Section 3.4. The NOMA and SWIPT equations used in Section 3.7 are shown in Section 3.5, while NOMA and SWIPT for MMC Networks scheme is outlined in Section 3.6. A comparison between the proposed scheme and traditional K-means based on the MATLAB simulation results is shown in Section 3.7. Section 3.8 concludes this chapter.

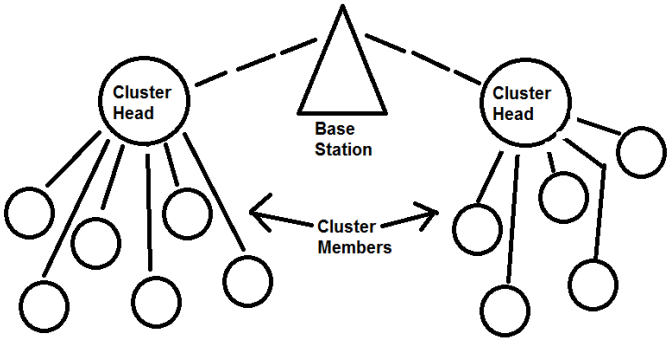


Figure 3.1: The devices within the M2M network are arranged into k clusters based on K-means Clustering. © 2018 IEEE, from [2].

3.2 System Model

As established in Subsection 2.5.2, a set of static d M2M devices (such as for the smart grid) arranged into a set of k clusters by using K-means Clustering, can be represented as $n \in \{1, 2, 3, \ldots, d\}$ and $m \in \{1, 2, 3, \ldots, k\}$, respectively. Each n-th device can then be referred individually by D_n, or referred to the m-th cluster it belongs to by $D_{n,m}$. The system model with an example of 2 clusters is shown in Figure 3.1. Within each cluster, NOMA is utilised, so that each device can share the same RB in LTE-A. In 5G, NOMA could potentially be applied to smaller frequency bands and smaller time slots (see Subsubsection 2.2.3.2).

The devices are arranged so that $|h_1|^2 > |h_2|^2 > \cdots > |h_d|^2$, where $|h_n|^2$ is the Rayleigh fading channel gain between the eNodeB and D_n. The power allocated to D_n, $P_n = \alpha_n P_T$, is $P_d > \cdots > P_2 > P_1$ (i.e. the strongest $|h_n|^2$, $|h_1|^2$, is allocated the lowest P_n, P_1). The strongest signal (weakest device's signal), $x_d(t)$, is decoded first and then subtracted from $y_n(t)$. The SIC process is done iteratively until the device has decoded its own signal,

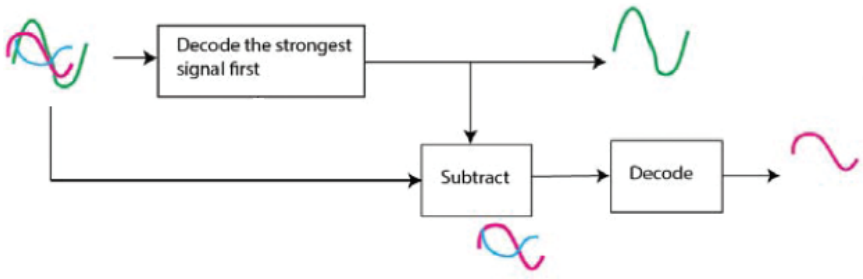


Figure 3.2: Signals are iteratively decoded and subtracted from the superimposed signal, $y_n(t)$, from the strongest signal (weakest device's signal) to the weakest signal [70].

$x_n(t)$, by treating the weaker devices' signals as noise, as shown in Figure 3.2 [51, 70].

3.3 Improved K-means Algorithm

An improved K-means algorithm termed K-means++, was proposed in [114], which reduces the convergence time and provides a better clustering solution than the traditional K-means algorithm, by modifying how the cluster centroid locations are initially chosen. K-means++ Clustering has an effectiveness of $\mathcal{O}(\log k)$-competitive to the optimal K-means clustering solution and a complexity of $\mathcal{O}(kn)$ [114]. The steps for K-means++ Clustering is given in **Algorithm 3.1**, where $d_{n',m}$ is the shortest Euclidean distance between point n' and the closest centroid already chosen.

Algorithm 3.1: K-means++ Clustering

1. Choose the first centroid location (C_{x1}, C_{y1}), uniformly at random from set χ.
2. Choose the next centroid location (C_{xm}, C_{ym}), $m = 2, 3, \ldots, k$, where $(C_{xm}, C_{ym}) = z \in \chi$, with the probability in Equation (3.1).
3. Repeat Step 2 for each m, until k centroids have been chosen.
4. Step 2–4 of **Algorithm 2.2** (i.e. the traditional K-means) is executed.

$$\frac{(d_{z,m})^2}{\sum_{n' \in \chi} (d_{n',m})^2} \tag{3.1}$$

The results in [114] over 20 simulation runs, show that K-means++ achieves a lower average SSE and a factor of 2 or better speed-up of the average running time, compared to the traditional K-means algorithm, for $k = \{10, 25, 50\}$. There were four different datasets used with $n = 10000$ in $\mathbb{R}^{15}$, $n = 1024$ in $\mathbb{R}^{10}$, $n = 494019$ in $\mathbb{R}^{35}$, and $n = 4601$ in $\mathbb{R}^{58}$, respectively, where $\mathbb{R}^r$ is the set of Real numbers in the r-dimensional vector space. A scalable K-means++ algorithm termed *K-means*$\|$, was proposed in [115].

3.4 Proposed Enhanced K-means Clustering with NOMA

Based on the 'User Pairing' concept in [53], it is advantageous to have at least one strong channel gain device within each cluster, to enhance the NOMA network sum throughput gain over OMA. Therefore, different from **Algorithm 3.1** and [114], it is proposed in **Algorithm 3.2** that the k strongest channel gain devices are excluded before clusters are initially formed (see Step 3), and then re-assigned as CHs to the appropriate cluster, in accordance with 'User Pairing' (see Step 5). This reassignment of devices also eliminates any size 1 clusters that could be formed when using the traditional K-means, K-means++, or in step 1 of **Algorithm 3.2**.

The 'silhouette value' of D_n, S_n, is one metric that can be used to determine if the clusters formed by using K-means, have well-matched devices. S_n is given in Equation (3.2), where a_n is the average distance between D_n and the other devices within the same cluster, and b_n is the minimum average distance between D_n to devices within the nearest cluster [116].

$$S_n = \frac{b_n - a_n}{\max(a_n, b_n)} \in [-1, 1] \tag{3.2}$$

The K-means algorithm is run for a range of k values, where D_n is well-matched to its own cluster if $S_n \approx 1$, while the number of clusters formed is non-optimal if $S_n \approx 0$ or < 0. The average S_n of the d number of devices is given in Equation (3.3), and will eventually peak when the optimal k value is found. This 'peak' is shown in Figure 3.3, where $k = 10$ is the optimal number of clusters for the accompanying data set [97].

$$S_{\text{avg}} = \frac{\sum_{n=1}^{d} S_n}{d} \tag{3.3}$$

The K-means algorithm is proposed to be run 25 times (based on page 554 of [97]), while using the optimal k value, with differing initial centroid locations. As shown in Equation (3.4), the lowest *sum of squared error* (**SSE**) — sum of the sum of squared Euclidean distances between devices and centroid in each cluster — is the best solution to the cluster formation problem, based on the number of times K-means is run [97]. Obviously, the K-means algorithm can be run for more than the proposed 25 times, to ensure a higher probability of finding the absolute lowest **SSE**. However, there is a trade-

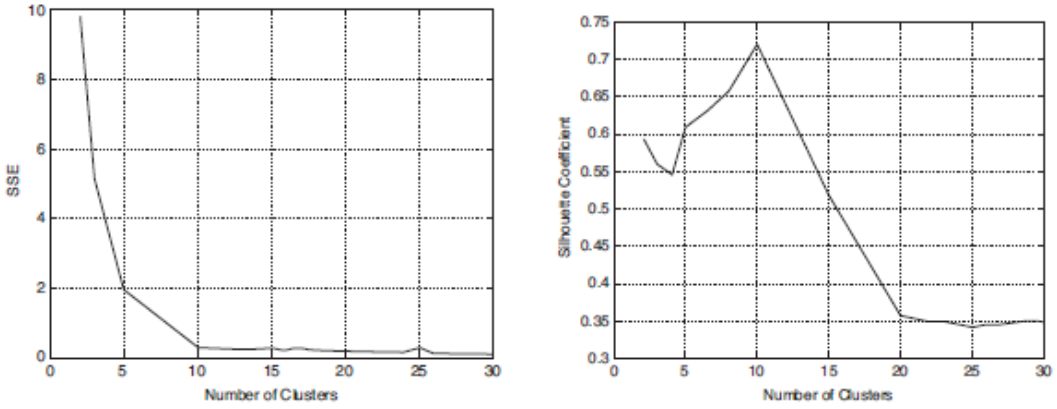


Figure 3.3: Left: SSE vs number of clusters, **Right**: S_{avg} vs number of clusters. © 2006. Reprinted from [97], by permission of Pearson Education, Inc., New York, USA.

off between obtaining a better solution to the cluster formation problem and the total running time of the algorithm [97].

$$\textbf{SSE} = \sum_{m=1}^{k} \sum_{N_{in} \in C_{im}} (d_{n,m})^2 \tag{3.4}$$

The **SSE** settles at the optimal $k = 10$ number of clusters for the accompanying data set, as shown in Figure 3.3 [97]. Note that $N_{in} = (N_{xn}, N_{yn})$, $C_{im} = (C_{xm}, C_{ym})$, and $d_{n,m}$, were defined earlier in Subsection 2.4.2, as the n-th device location, m-th centroid location, and the Euclidean distance between the n-th device and m-th centroid, respectively.

The steps for the proposed enhanced K-means algorithm is given in **Algorithm 3.2** (where K-means++ in **Algorithm 3.1** is used to select the centroid locations in a specific way, rather than randomly, followed by step 2-4 of the traditional K-means algorithm in **Algorithm 2.2**), where $\lceil \cdot \rceil$ is the ceiling function. Clusters are expected to be of $>$ size 1, to minimise the SIC complexity and to also maintain the NOMA network sum throughput advantage over OMA. The $\lceil d/2 \rceil$ factor in step 1, aims to ensure that clusters formed may all be $>$ size 1, while step 5 guarantees that all clusters are $>$ size 1. S_{avg} from Equation (3.3) is used to find the optimal k number of clusters (i.e. the k where $S_{\text{avg}} \rightarrow 1$). This signifies that the devices are well-matched to their clusters, but the NOMA network sum throughput is not necessarily maximised until Steps 3 and 5 are applied.

Algorithm 3.2: Enhanced K-means Clustering for MMC Networks

1. The K-means algorithm is run for $m = \{1, 2, \ldots, \lceil d/2 \rceil\}$. Compute S_{avg} for each m, where $S_{\text{avg}}(m) = \{S_{\text{avg}}(m = 1), S_{\text{avg}}(m = 2), \ldots, S_{\text{avg}}(m = \lceil d/2 \rceil)\}$.
2. Let $k = \underset{m}{\text{argmax}}\, S_{\text{avg}}(m)$.
3. The k devices with the strongest channel gains are excluded from the network.
4. **Algorithm 3.1** is executed.
5. The excluded k number of devices are assigned as CHs to the appropriate cluster.

3.5 NOMA and SWIPT

The SINR at D_n to decode $x_n(t)$ (assuming perfect SIC) can be found by Equation (3.5), where B is the transmission bandwidth. Note that $x_n(t)$, P_n, $|h_n|^2$ and N_0 were defined earlier in Subsection 2.5.2, as the individual information conveying OFDM waveform, the DL power allocated to D_n, the Rayleigh fading channel gain between D_n and the eNodeB, and the spectral noise power density of the AWGN channel, respectively. It was assumed earlier in Section 3.2, that devices are ordered as $|h_1|^2 > |h_2|^2 > \cdots > |h_d|^2$.

$$\text{SINR}_n = \frac{P_n |h_n|^2}{N_0 B + \sum_{i=1}^{n-1} P_i |h_n|^2}, \quad [\text{dB}] \tag{3.5}$$

For an example 2-device cluster (i.e. UE1/UE2 device pair), the SINR at UE1 while decoding UE2's signal, $x_2(t)$, can be expressed as $\text{SINR}_{1\to2}$, resulting in $P_n = P_2$, with $|h_1|^2$, as shown in Equation (3.6).

$$\text{SINR}_{1\to2} = \frac{P_2 |h_1|^2}{N_0 B + P_1 |h_1|^2} \tag{3.6}$$

The resulting throughput for D_n when utilising DL NOMA and OMA, can be found by Equations (3.7) and (3.8) below, respectively.

$$\eta_n = B\log_2\left(1 + \frac{P_n |h_n|^2}{N_0 B + \sum_{i=1}^{n-1} P_i |h_n|^2}\right), \quad [\text{bps}] \tag{3.7}$$

$$\eta_n = \frac{B}{d}\log_2\left(1 + \frac{P_n\left|h_n\right|^2}{N_0\frac{B}{d}}\right) \tag{3.8}$$

The network sum throughput (i.e. the total throughput for a network of d devices) is shown in Equation (3.9), where the spectral efficiency is (η_T / B).

$$\eta_T = \sum_{n=1}^{d}\eta_n \tag{3.9}$$

When the PS SWIPT EH strategy (i.e. the received RF signal is split for EH and information decoding) is applied to D_n, for $n = 1, 2, \ldots, (d-1)$, then Equation (3.7) can be rewritten as shown in Equation (3.10), where $0 \leq P_s \leq 1$ is the portion of the received normalised signal used for EH [58]. SWIPT is not required for the weakest channel gain device, as it is able to decode its own signal directly, by treating the other signals as noise.

$$\eta_n = B\log_2\left(1 + \frac{P_n(1-P_s)\left|h_n\right|^2}{N_0B + \sum_{i=1}^{n-1}P_i(1-P_s)\left|h_n\right|^2}\right) \tag{3.10}$$

For optimal EE, P_s must be large enough to satisfy Equation (3.11), where $0 \leq \zeta \leq 1$ is the EH efficiency, P_T is the DL network power budget, and P_{SIC} is the SIC power consumption. It is assumed that $\zeta = 0.5$, as the EH efficiency is shown to be linear in the low power region [58].

$$P_s\zeta\left|h_n\right|^2 P_T \geq P_{\text{SIC}} \tag{3.11}$$

Assuming that P_{SIC} is proportional to the SIC complexity of $\mathcal{O}(L^3)$ as mentioned earlier in Section 2.6, then P_{SIC} increases by $\mathcal{O}(L^3)$ / $\mathcal{O}\left[(L-1)^3\right]$, for $L = 3, 4, \ldots, \lceil d/2 \rceil$. There is a $\lceil d/2 \rceil$ factor, since the complexity of SIC is not relevant for clusters of size 1. We assume that $P_{\text{SIC}} = 80$ mW for a 2-device cluster as in [58].

3.6 NOMA for MMC Networks

Devices within each cluster will be allocated power based on their Rayleigh fading channel gains, $\left|h_n\right|^2$, as mentioned earlier. The CH of each cluster will be responsible for providing information to the eNodeB of the power requirements within its cluster.

The power allocated to D_n, can be written as $P_n = \alpha_n P_T$, where α_n is the power allocation coefficient for D_n, and $\sum_{n=1}^{d}\alpha_n = 1$. This leads us to an important question

of, "**How should P_T be allocated to the devices within the network, to maximise η_T while under various constraints?**". If n is large, then it is impractical to consider the many different combinations of allocating power, which leads to the formulation of a convex optimisation problem [76], as shown in Equation (3.12). η_n is given in Equation (3.7).

$$\begin{aligned} \underset{\alpha_n}{\text{maximise}} \quad & \eta_n \\ \text{subject to:} \quad & \sum_{n=1}^{d} P_n \leq P_T \\ & \sum_{n=1}^{d} \alpha_n = 1 \\ P_n \geq 0, \quad & n = 1, \ldots, d \end{aligned} \tag{3.12}$$

In [76], KKT optimality conditions were used to derive closed-form solutions for optimal power allocation (see Subsection 2.5.7), based on an optimisation problem like Equation (3.12). The large number (i.e. 2^{2m} for a cluster of size m) of Lagrange multiplier combinations required to be checked to satisfy the KKT conditions, were reduced to a general 2^{m-1} power allocation equations, since $P_n \geq 0$. Each set of power allocation equations has a corresponding set of necessary conditions that must be satisfied, as mentioned in Subsection 2.5.7. The power allocation equations and corresponding necessary conditions consist of $\varphi_i = 2^{\beta}$, where $\beta = (R_i \,/\, \omega B_{\text{RB}})$. R_i is the minimum rate requirement, ω is the number of RBs allocated to a cluster, and $B_{\text{RB}} = 180$ kHz is the bandwidth of a RB. To make a fair comparison between NOMA and OMA, ω is equal to the size of the particular cluster. φ_i influences which set of power allocation equations is selected for each cluster. Therefore, φ_i also influences Equations (3.7) and (3.9), because of R_i. The DL power budget for each cluster is $P_t = (\omega P_T \,/\, B_{\text{tot}})$, where B_{tot} is the number of available RBs.

As an example, there are two possible power allocations that can be used for a 2-device NOMA cluster (i.e. $n = 1, 2$), as shown in Equations (3.13) and (3.14). Note that $\gamma_n = \sqrt{|h_n|^2 \,/\, N_0 B_{\text{RB}}}$, is the normalised channel gain [76].

$$P_1 = \frac{P_t}{\varphi_2} - \frac{\omega(\varphi_2 - 1)}{\varphi_2 \gamma_2}, \quad P_2 = \frac{P_t(\varphi_2 - 1)}{\varphi_2} + \frac{\omega(\varphi_2 - 1)}{\varphi_2 \gamma_2} \tag{3.13}$$

$$P_1 = \frac{P_t}{2} - \frac{P_{\text{tol}}}{2\gamma_1}, \quad P_2 = \frac{P_t}{2} + \frac{P_{\text{tol}}}{2\gamma_1} \tag{3.14}$$

The corresponding necessary conditions for Equations (3.13) and (3.14), are shown in Equations (3.15) and (3.16) respectively, where only one set of power allocations can be used at one time, depending on the value of R_i. If R_i is too high, then none of the necessary conditions are satisfied, meaning that R_i cannot be met for all cluster devices.

$$P_n\gamma_n - \left(\varphi_1 - 1\right)\left(\sum_{j=1}^{n-1} P_j\gamma_n + \omega\right) > 0,\ n = 1 \tag{3.15a}$$

$$\left(P_n - \sum_{j=1}^{n-1} P_j\right)\gamma_{n-1} - P_{\text{tol}} > 0,\ n = 2 \tag{3.15b}$$

$$P_n\gamma_n - \left(\varphi_1 - 1\right)\left(\sum_{j=1}^{n-1} P_j\gamma_n + \omega\right) > 0,\ n = 1, 2 \tag{3.16}$$

The steps for the proposed NOMA algorithm for MMC clusters are given in **Algorithm 3.3**. Note that SWIPT in step 4 of **Algorithm 3.3** is an extension of the proposed **Algorithm 3.2** and the optimal power allocation in step 3. Thus, the resulting solution from the optimisation problem of Equation (3.12)[1] is not applicable to SWIPT.

Algorithm 3.3: NOMA for MMC Network

1. Calculate h_n for each device [as in Equation (2.9)], based on their distance d_n from the eNodeB.
2. Within each cluster, sort the devices in descending $|h_n|^2$, so that P_t can be allocated according to Table 1 in [76].
3. Determine which set of power allocation equations to use for a cluster, based on size and the corresponding set of satisfied necessary conditions.
4. Calculate the required P_s for each device utilising SIC within a cluster, to satisfy Equation (3.11).

[1]Equation (3.12) can easily be extended to include SWIPT (i.e. results shown in Figures 3.9a–3.11), but the enhanced K-means clustering algorithm developed in **Algorithm 3.2** (i.e. results shown in Figure 3.12), is the main outcome of this thesis.

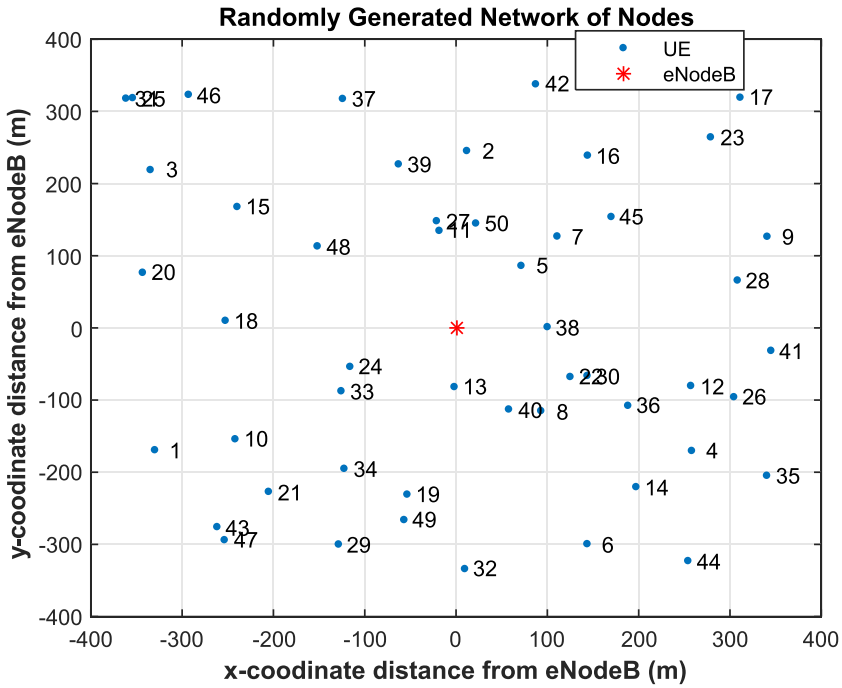


Figure 3.4: An example network of 50 devices, labelled from D_1 to D_{50}.

3.7 Results and Discussion

An example network of 50 M2M devices (as shown in Figure 3.4), was generated in MATLAB from a spatial Poisson point process with intensity $\lambda = 6$, and then the accompanying x- and y-coordinates were randomly chosen from a continuous uniform distribution.

3.7.1 K-means Clustering

The average silhouette value for $k = \{1, 2, \ldots, 25\}$ was first computed (see Step 1 in **Algorithm 3.2**), with the highest being the optimal k value. An example of this 'peak' in the average silhouette value, to signify the optimal k value, was shown in Figure 3.3. The k strongest channel gain devices were excluded from the network in Figure 3.4. The K-means algorithm was run 25 times for the remaining network with $k = 20$ (as shown in Figure 3.5), with differing initial centroid positions, for a higher probability of finding the best clustering solution. Each of the excluded k strongest channel gain devices were assigned to the appropriate cluster as CHs to enhance the network sum throughput, according to 'User Pairing', and to resolve any size 1 clusters that may have been formed in step 1 of **Algorithm 3.2**.

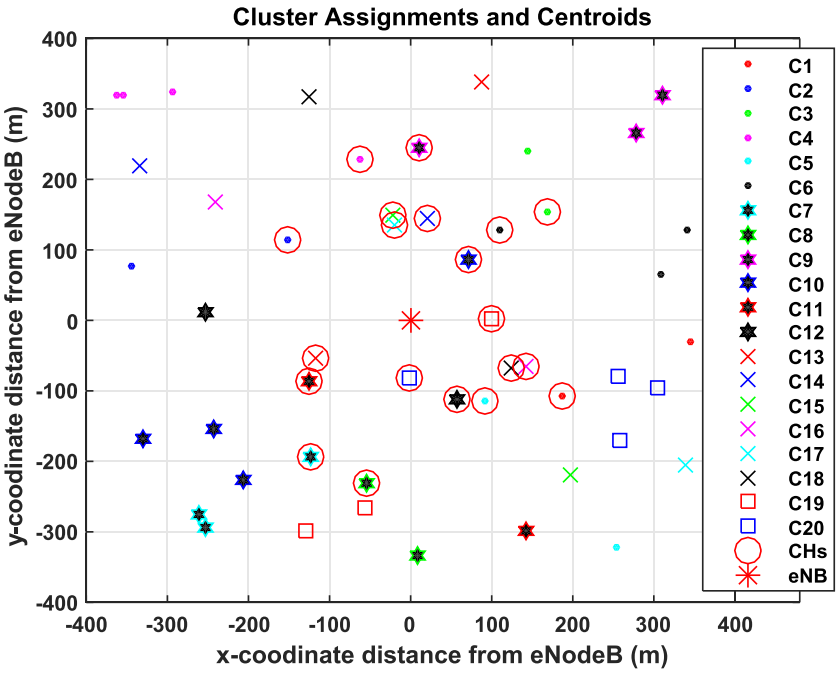


Figure 3.5: The example network from Figure 3.4, split into $k = 20$ clusters by using the enhanced K-means algorithm.

3.7.2 NOMA

The simulation parameters given in **Table 3.1**[2], were used for both NOMA and OMA. X from Equation (2.9) was generated, and then $|h_n|$ was computed for every device within the network. For each cluster, $P_t = (\omega P_T \ / \ 50)$, where $P_T = 46$ dBm $= 39.8107$ W, for an eNodeB in LTE-A [117]. The path loss exponent of $\alpha = 3.76$, is a standard value used in LTE-A simulations, for a suburban scenario [117], which was used to generate Figures 3.6, 3.7 and 3.12.

Using Equation (3.9), the network sum throughput was determined while utilising NOMA and OMA respectively, by performing the summation of Equations (3.7) and (3.8). The resulting sum throughput for each cluster, is shown in Figure 3.6. It can be observed that the sum throughput of each NOMA cluster is higher than the corresponding OMA cluster. However, the NOMA sum throughput gain over OMA diminishes as the device channel gains within a cluster approach the same value [54] (e.g. C3 and C9 from Figure 3.5).

[2]Note that the simulation uses the normalised channel gain (i.e. $\gamma_n = \sqrt{|h_n|^2} \ / \ N_0 B_{\mathrm{RB}}$) instead of N_0, based on the simulation results in [76], to generate Figures 3.9a–3.11.

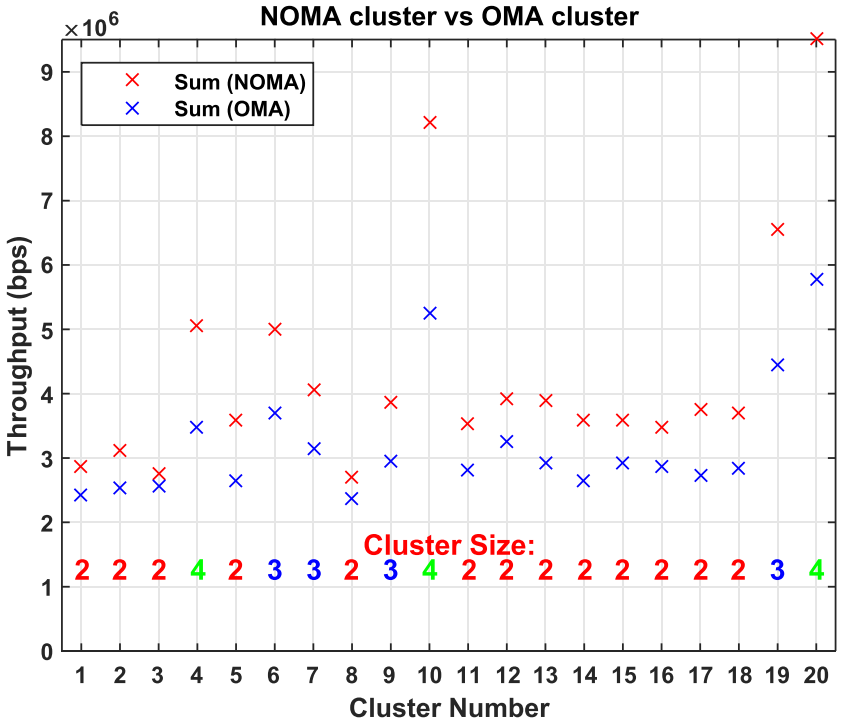


Figure 3.6: Sum throughput comparison between NOMA and OMA clusters formed using the enhanced K-means algorithm, for $R_i = 100$ kbps.

Table 3.1: MATLAB DL Simulation Parameters

Parameters	**Setup**
Bandwidth, W	10 MHz
Number of available RBs, B_{tot}	50
Transmit Antennas	1
Receive Antennas	1
DL Transmit Power Budget, P_T	46 dBm [117]
Minimum SIC threshold, P_{tol}	10 dBm [76]
Path loss exponent, α	3.76 [117]
ω	2–4 [76]
Minimum rate requirement, R_i	100 kbps
Energy harvesting efficiency, ζ	0.5 [58]
SIC power consumption, P_{SIC}	80 mW and 270 mW [58]

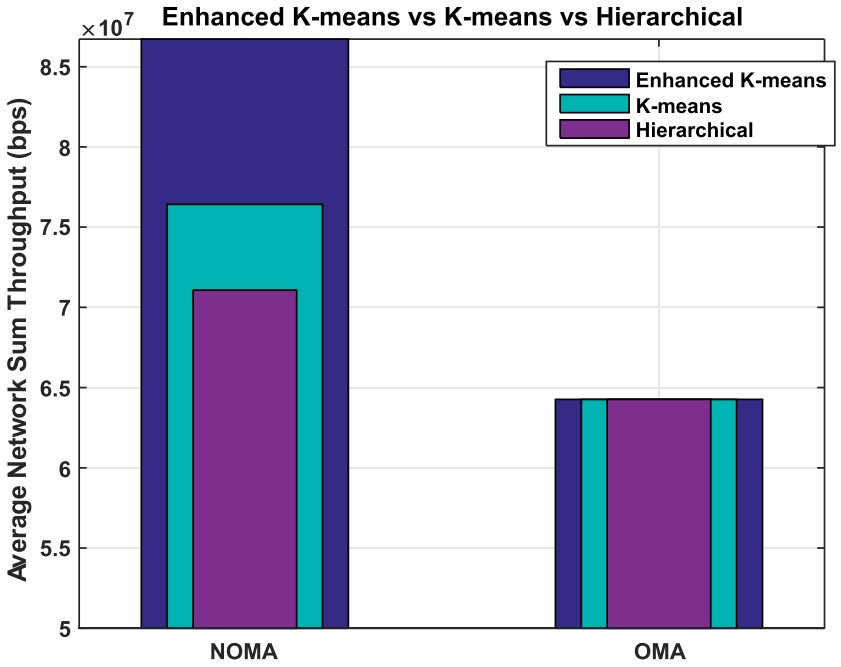


Figure 3.7: NOMA vs OMA network sum throughput (on average) comparison between the enhanced K-means, traditional K-means and Hierarchical, for $R_i = 100$ kbps.

3.7.3 Enhanced K-means vs Traditional K-means

The NOMA network sum throughput for the proposed enhanced K-means was compared with the traditional K-means and Hierarchical Clustering (as verification of why K-means is preferable over Hierarchical, as stated in Section 2.4) as shown in Figure 3.7. The simulation was run the proposed 25 times for both schemes and Hierarchical Clustering, and the results were averaged out to mitigate the differing of the 'best clustering solution' from each simulation run. The averaged results are summarised as follows:

- Hierarchical was around 10.59% higher (maximum of 10.74% and minimum of 10.53%) than OMA. The traditional K-means was around 18.92% higher (maximum of 19.04% and minimum of 18.83%) than OMA.
- The enhanced K-means NOMA network sum throughput was around 34.94% higher (maximum of 35.03% and minimum of 34.86%) than OMA. This represents an improvement of around 13.47% (on average) over the traditional K-means and 22.00% (on average) over Hierarchical, when using the proposed enhanced K-means.

The proposed scheme achieves a higher percentage gain over OMA compared to tradi-

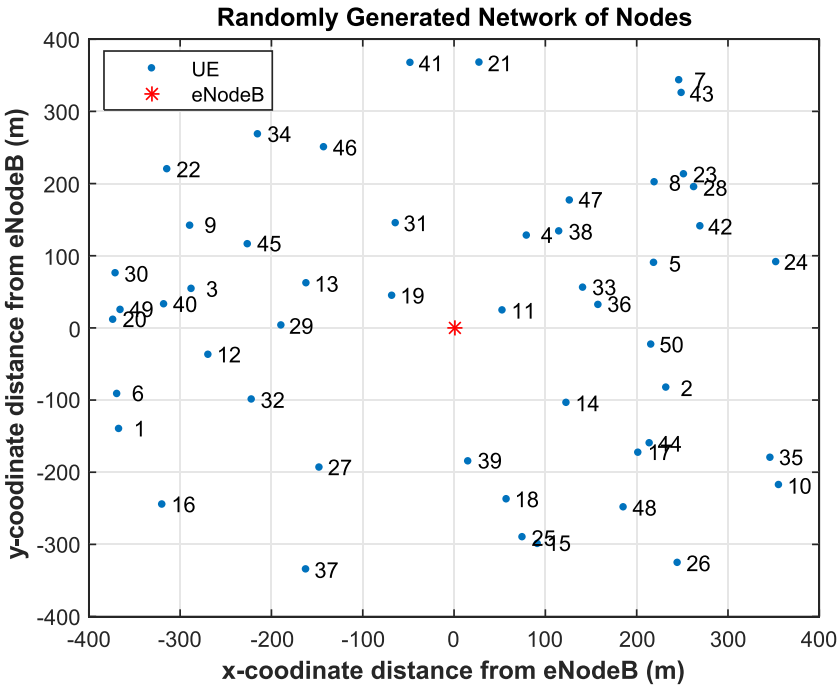


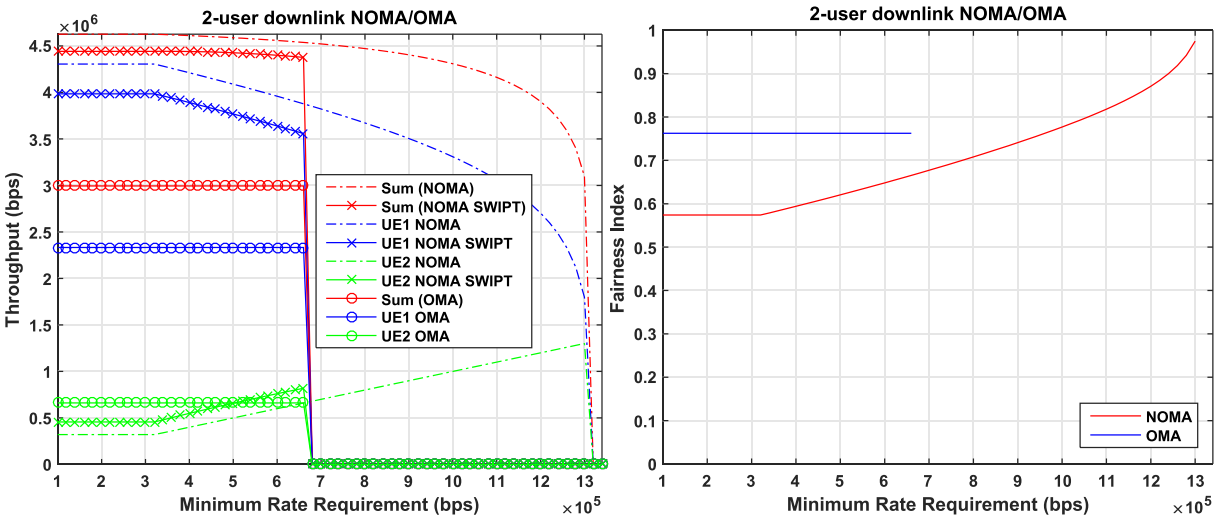
Figure 3.8: An example 2nd network of 50 devices, labelled from D_1 to D_{50}.

tional K-means, because 'User Pairing' was used to enhance the network sum throughput. Referring to Figure 3.4, the traditional K-means would most likely assign D_8, D_{13}, D_{22}, D_{30}, D_{38} and D_{40} into one cluster. However, each of these devices are one of the k strongest channel gain devices, which means they would instead be assigned as CHs to separate clusters (as shown and labelled accordingly in Figure 3.5) when the proposed scheme is utilised.

As another example, the proposed scheme was run for the network in Figure 3.8, with a different distribution of device locations in relation to the eNodeB. The averaged results are summarised below:

- The enhanced K-means NOMA network sum throughput was 30.43% higher than OMA, where the network sum throughput gain is 4.67% lower (essentially similar) than the network in Figure 3.4.

- The OMA network sum throughput is 3.87% lower than the network in Figure 3.4, indicating a lower average channel gain for devices.



(a) (b) Fairness Index Comparison

Figure 3.9: NOMA vs OMA at varying R_i with $|h_1|^2 = (40N_0B_{\text{RB}})^2$ dB and $|h_2|^2 = (11.7N_0B_{\text{RB}})^2$ dB. **NOTE: the throughput is set to 0 bps (and fairness index is undefined) when the system is in 'outage' or when DL SIC is not feasible.**

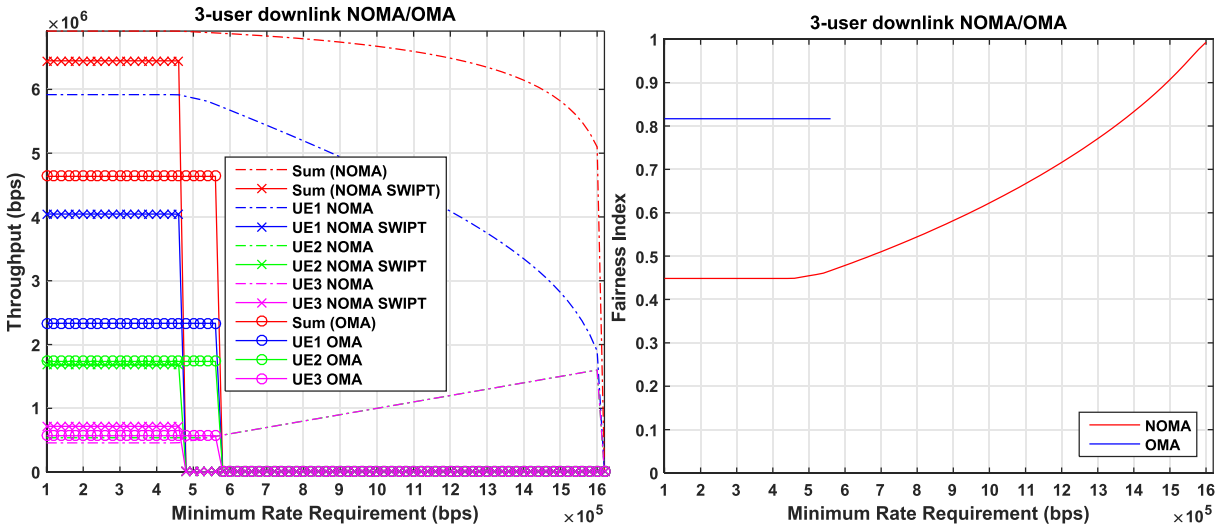
3.7.4 Varying Minimum Rate Requirements

As mentioned in Section 3.6, only one set of power allocation equations can be used at one time. The selected set of power allocation equations is based on the value of R_i and also how different the $|h_n|^2$ of devices are within a particular cluster.

3.7.4.1 2-Device and 3-Device Cluster Example with SWIPT

A 2-device cluster with $|h_1|^2 = (40N_0B_{\text{RB}})^2$ dB and $|h_2|^2 = (11.7N_0B_{\text{RB}})^2$ dB was considered. $R_i = \{100, 120, \ldots, 1340\}$ kbps is set as shown in Figure 3.9a. It can be observed:

- The NOMA sum throughput is constant for 100 kbps $\leq R_i \leq$ 320 kbps, because the power allocated to both devices are similar according to Equation (3.14).
- The sum throughput starts to decrease when $R_i >$ 320 kbps, since Equation (3.13) is used instead to apply power control to the weaker device. This also leads to an improvement in the 'fairness index' as shown in Figure 3.9b [see Equation (2.12)] between both devices by allocating an increasing portion of P_t to the weaker device so that R_i can be met, but this reduces the throughput of the strong device.



(a) **(b)** Fairness Index Comparison

Figure 3.10: NOMA vs OMA at varying R_i with $|h_1|^2 = (40N_0B_{\text{RB}})^2$ dB, $|h_2|^2 = (30N_0B_{\text{RB}})^2$ dB and $|h_3|^2 = (10N_0B_{\text{RB}})^2$ dB.

- From $R_i > 1300$ kbps for NOMA, and $R_i > 660$ kbps for OMA, Equation (3.13) and Equation (3.14) can no longer be satisfied as R_i cannot be met for both devices, as in the system is deemed in 'outage'.

Using Equation (3.11) with $P_s = 0.5$ and $P_{\text{SIC}} = 80$ mW, Figure 3.9a shows that:

- DL SIC is feasible[3] for 100 kbps $\leq R_i < 700$ kbps.
- There is a slight decrease in the NOMA sum throughput due to P_1 being split between information decoding and EH. When varying P_s with $R_i = \{100, 120, \ldots, 1000\}$ kbps, it was found that NOMA is superior to OMA at all network conditions if $P_s \geq 0.46$.

Extending to a 3-device cluster, Figure 3.10a shows that DL SIC is feasible for 100 kbps $\leq R_i < 480$ kbps, if $P_s \geq 0.91$ and ≥ 0.27 for the strong device and middle device respectively, and with $P_{\text{SIC}} = 270$ mW and 80 mW, respectively. **This is a promising**

[3]The energy harvested via PS SWIPT was shown to be able to power DL SIC in [58], for a 2-device cluster. For the sample 50-device network in Figure 3.4, based on Figure 3.9a, the curves in Figure 3.12 would shift towards the lower left corner to facilitate DL SIC, which is not much of a concern since MMC devices are expected to be transferring small amounts of data.

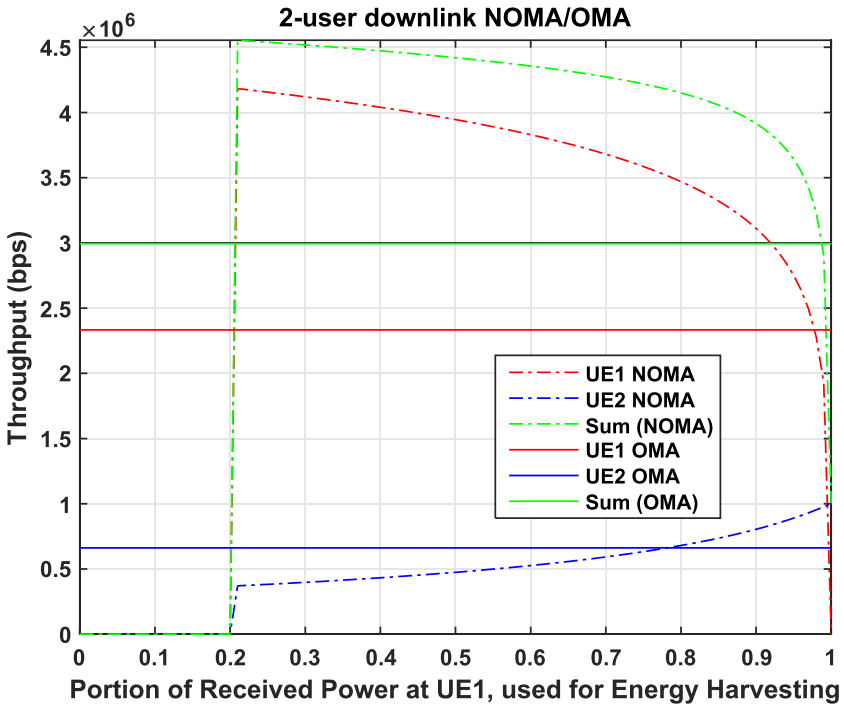
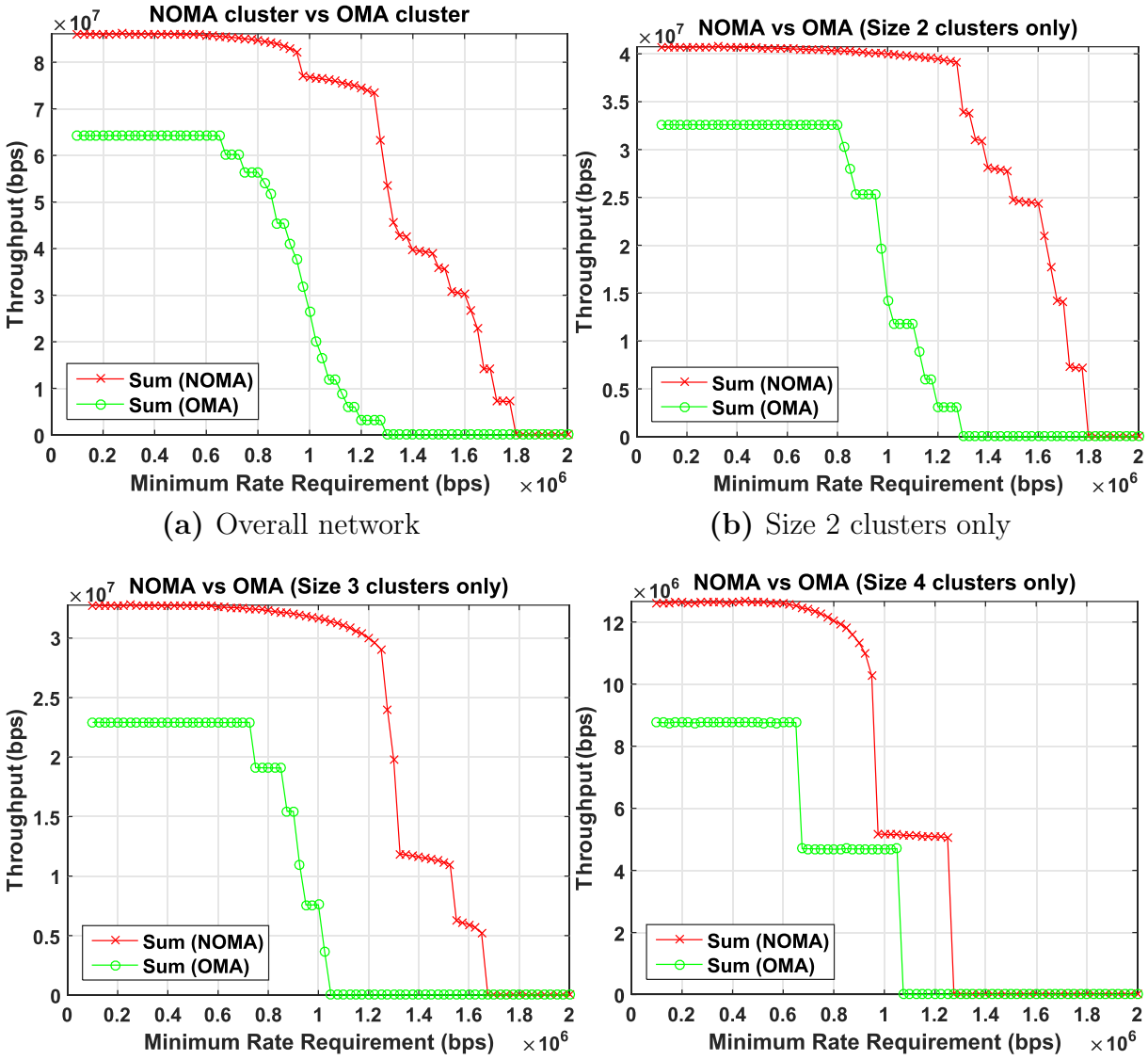


Figure 3.11: NOMA vs OMA at varying P_s for SWIPT, for $R_i = 100$ kbps. **NOTE: the throughput is set to 0 bps when DL SIC is not feasible.**

result for utilising DL SIC for cluster sizes of larger than 2, as to the best of our knowledge, other papers dealing with SWIPT in DL NOMA have only considered clusters of size 2. In Figure 3.11, we set $P_s = \{0, 0.01, \ldots, 1\}$ for a 2-user cluster. At $R_i = 100$ kbps, DL SIC is feasible for $0.2 < P_s < 1$, where NOMA is superior to OMA within that region.

3.7.4.2 Enhanced K-means

The proposed enhanced K-means algorithm was run at $R_i = \{100, 120, \ldots, 2000\}$ kbps, as shown in Figure 3.12a, based on the network in Figure 3.4. It can be observed that the enhanced K-means network sum throughput is higher than OMA for all R_i. The network sum throughput starts to decrease as expected when a new set of power allocation equations is used. However, it is obvious that the necessary conditions for 4 to 2-device clusters respectively, can no longer be satisfied as R_i is increased towards 2000 kbps.



(a) Overall network

(b) Size 2 clusters only

(c) Size 3 clusters only

(d) Size 4 clusters only

Figure 3.12: NOMA vs OMA network sum throughput with varying R_i. **NOTE: the steep drops in sum throughput is due to a cluster or clusters being in 'outage'.**

3.7.4.3 Different Sized Clusters

The 4-device clusters in Figure 3.4 begin to experience outage from $R_i > 950$ kbps, while the 3-device clusters from $R_i > 1250$ kbps. Therefore, it is easier to satisfy the necessary conditions for smaller cluster sizes, meaning that larger networks are limited to lower R_i. This can be seen in Figure 3.12b–3.12d for 2 to 4-device clusters respectively from Figure 3.4. Refer to Figure 3.6 for the cluster sizes for the network in Figure 3.4.

This R_i limitation would not be such a problem, as these MMC devices are expected to be transferring small amounts of data. It is much more important to be able to simultaneously accommodate the billions of devices within an MMC network. The enhanced K-means algorithm achieves this as it avoids the possibility of an overload and access problem by placing devices into clusters and utilising NOMA. However, when $R_i > 950$ kbps, some of the clusters will be in outage for the Figure 3.4 network. Having smaller cluster sizes would increase the R_i threshold for when the system is in outage for NOMA, and there is a particular cluster size (which is 2 to 4 based on our results) where NOMA is superior to OMA at all network conditions.

3.8 Conclusion

In this chapter an enhanced K-means algorithm with DL NOMA and SWIPT was proposed, to fulfil the requirements of MMC devices (such as sensor nodes), which have the QoS requirements of massive connectivity, small packets, low data rate, high EE, and low energy consumption. MMC devices will likely suffer from the *overload and access problem*, if the OMA technique is used. Clustering was used to reduce the feedback and control signalling overhead, while NOMA was used to avoid the *overload and access problem*.

The proposed scheme involved determining the optimal value of k clusters by using the 'silhouette value' metric (as in Section 3.4), and then the K-means clustering algorithm was applied to the network. The k strongest channel gain devices were excluded from the network and were then reassigned to the appropriate cluster to ensure that each cluster had a strong channel gain device, in accordance with 'User Pairing'. Power was optimally allocated to each device within a cluster based on a set of equations and corresponding necessary conditions (as in Section 3.6).

The earlier mentioned clustering was also used to reduce the power consumption of the NOMA SIC process compared to an un-clustered MMC network. In order to address the EE of the SIC process, the PS SWIPT EH technique was used to mitigate the power consumption of the SIC process.

The proposed scheme was shown in Figure 3.6 to have a higher NOMA cluster sum throughput for the $k = 20$ clusters formed, at a minimum rate requirement of $R_i = 100$ kHz. It was also shown in Figure 3.7 to have a 34.94% higher average NOMA network sum throughput (over 25 simulation runs) over OMA, a 13.47% improvement over traditional

K-means and a 22.00% improvement over Hierarchical, at a minimum rate requirement of 100 kHz. For 2-device clusters, it was shown in Figure 3.9a, that DL SIC is feasible for 100 kbps $\leq R_i <$ 700 kbps, with a slight decrease in the NOMA sum throughput. NOMA had a superior sum throughput when the portion of the received normalised signal used for EH, P_s, was ≥ 0.46. It was also shown in Figure 3.11, that DL SIC is feasible for $0.2 < P_s < 1$. Extending to 3-device clusters, it was shown in Figure 3.10a, that DL SIC is feasible for 100 kbps $\leq R_i <$ 480 kbps. The proposed scheme was shown in Figure 3.12 to have a higher network sum throughput improvement over OMA compared to the traditional K-means with 100 kbps $\leq R_i \leq$ 2000 kbps. Applying PS SWIPT showed promising results in increasing the EE by mitigating the power consumption of SIC.

Chapter 4

Uplink NOMA with SWIPT

4.1 Introduction

4.1.1 Motivations

In this chapter, an enhanced K-means algorithm with UL NOMA and SWIPT is proposed [1], to fulfil the requirements of MMC devices (such as sensor nodes) and address some of the challenges mentioned at the beginning of Chapter 3. The TS SWIPT EH technique (mentioned in Section 2.6) is adopted to mitigate the power consumption of UL transmission and also for higher EE. The complexity (and therefore power consumption) of the SIC process is not as relevant in the UL, as it was in the DL, since the eNodeB has a much larger energy reserve compared to MMC/IoT devices.

4.1.2 Solution

The same clustering procedure was followed as in the DL (see Chapter 3 on page 53), but with an additional step in **Algorithm 4.1** to resolve the throughput problems observed in the UL. There is a subtle difference between the UL and DL (see Subsection 4.6.1 for more explanation), where the SIC process cannot easily decode the intended signals if there is too small of a difference between the channel gains of the 2 weakest devices and the second strongest device within a 4-device cluster.

This chapter is organised as follows: Section 4.2 establishes the system model. The proposed enhanced K-means with NOMA scheme is presented in Section 4.3. The NOMA and SWIPT equations used in Section 4.6 are shown in Section 4.4, while NOMA for

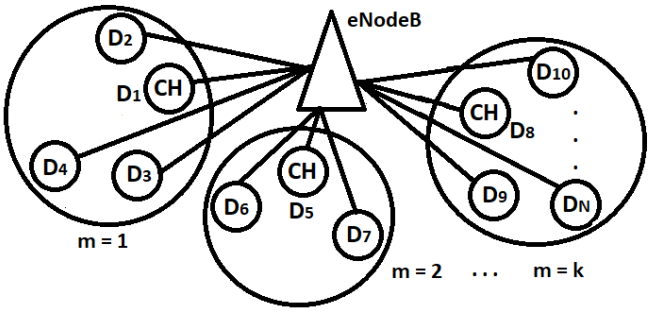


Figure 4.1: System model of n M2M devices arranged into m clusters. K-means Clustering is used as the basis for the proposed scheme.

MMC Networks is outlined in Section 4.5. A comparison between the proposed scheme and traditional K-means based on the MATLAB simulation results is shown Section 4.6. The combined DL/UL solution is given in Section 4.7, while Section 4.8 concludes this chapter.

4.2 System Model

An example system model of d M2M devices arranged into k clusters, by the use of K-means Clustering, was shown in Figure 4.1. These sets of M2M devices and clusters can be represented as $n \in \{1, 2, 3, \ldots, d\}$ and $m \in \{1, 2, 3, \ldots, k\}$, respectively. Each device can then be referred individually by D_n, or referred to the m-th cluster it belongs to by $D_{n,m}$. NOMA is used for intra-cluster communication, so that each device can share the same RB in LTE-A. In 5G, NOMA could potentially be applied to smaller frequency bands and smaller time slots (see Subsubsection 2.2.3.2).

The devices are arranged in decreasing $|h_n|^2$, as in the DL. The UL transmit power of D_n is $P_1 \geq P_2 \geq ... \geq P_d$. The strongest signal (strongest device's signal), $x_1(t)$, is decoded first and then subtracted from $y(t)$[1]. The SIC process is done iteratively until the eNodeB has decoded the signal of interest, $x_n(t)$, as shown in Figure 4.2.

[1]As mentioned in Subsection 2.5.2, this is the most optimal decoding order, since devices transmit at similar power levels and therefore the signal strength is determined by the devices' channel gains.

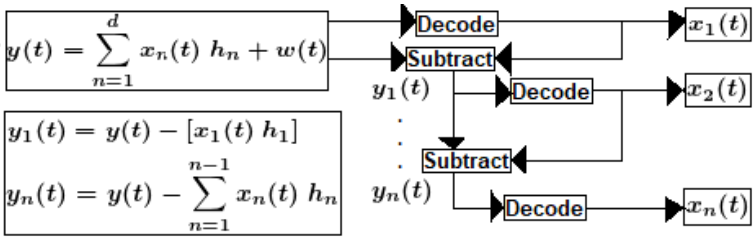


Figure 4.2: Signals are iteratively decoded and subtracted from $y(t)$, from the strongest channel device, $x_1(t)$, to the signal of interest, $x_n(t)$.

4.3 Proposed Enhanced K-means Clustering with NOMA

Based on the 'User Pairing' concept in [53], it is advantageous to have at least one strong channel gain device within each cluster. The k strongest channel gain devices are excluded when the clusters are initially formed as in the DL, and then re-assigned as CHs to the appropriate cluster, in accordance with 'User Pairing' and to eliminate any clusters of size 1 formed when using the traditional K-means or K-means++.

The K-means algorithm is run for a range of k values, and the 'silhouette value', which was shown in Equation (3.2), is computed for D_n. When the average silhouette value of the devices, S_{avg}, has 'peaked' (see Section 3.4), then the optimal k value has been found. The K-means algorithm is then proposed to be run 25 times (based on page 554 of [97]) with differing initial centroid locations, with the lowest **SSE** (see Section 3.4) being the best solution to the cluster formation problem, based on the 25 simulation runs.

The steps for the proposed enhanced K-means algorithm are given in **Algorithm 4.1**, where relevance of the additional Step 3 will be addressed in Subsection 4.6.1. Clusters are expected to be of $>$ size 1 **AND** $<$ size 5, to minimise the SIC complexity (and therefore power consumption), to maintain the NOMA network sum throughput advantage over OMA and to minimise the effects of SIC decoding error propagation (see Section 1.3.2). The $\lceil d/2 \rceil$ and $\lceil d/4 \rceil$ factors in step 1 (slightly modified from **Algorithm 3.2**), guarantees that clusters formed may all be $>$ size 1 and may all be $<$ size 4. Different from **Algorithm 3.2**, step 3 guarantees that all size 4 clusters are reduced to size 3 clusters.

Algorithm 4.1: Enhanced K-means Clustering for MMC Networks

1. Run the K-means algorithm for $m = \{\lceil d/4 \rceil, \lceil d/4 \rceil + 1, \ldots, \lceil d/2 \rceil\}$. Compute S_{avg} for each m, according to Equation (3.2), where $S_{\text{avg(m)}} = \{S_{\text{avg}}(m = \lceil d/4 \rceil), S_{\text{avg}}(m = \lceil d/4 \rceil + 1), \ldots, S_{\text{avg}}(m = \lceil d/2 \rceil)\}$.
2. Steps 2–5 of **Algorithm 3.2** are executed.
3. Re-assign the second strongest channel device in every 4-device cluster to the closest 2-device cluster.

4.4 NOMA and SWIPT

The SINR of D_n (assuming perfect SIC) can be found by Equation (4.1). Note that B, P_n, $|h_n|$ and N_0, were defined earlier in Section 3.5.

$$\text{SINR}_n = \frac{P_n |h_n|^2}{N_0 B + \sum_{i=n+1}^{d} P_i |h_i|^2}, \quad \text{[dB]} \tag{4.1}$$

The resulting throughput for each device when utilising UL NOMA and OMA, can be found by Equation (4.2) and (4.3) below, respectively.

$$\eta_n = B \log_2 \left(1 + \frac{P_n |h_n|^2}{N_0 B + \sum_{i=n+1}^{d} P_i |h_i|^2} \right), \quad \text{[bps]} \tag{4.2}$$

$$\eta_n = \frac{B}{d} \log_2 \left(1 + \frac{P_n |h_n|^2}{N_0 \frac{B}{d}} \right) \tag{4.3}$$

The network sum throughput (i.e. the total throughput for a network of d devices) is shown in Equation (4.4).

$$\eta_T = \sum_{n=1}^{d} \eta_n \tag{4.4}$$

Applying the TS SWIPT EH strategy to D_n (i.e. the transmission time slot is divided for EH and data transmission), Equation (4.2) can be rewritten as in Equation (4.5), where $0 \leq T_s \leq 1$ is the portion of the normalised transmission time used for EH, and $P'_n = E'_n \big(1 - T_s\big)^{-1}$ is the TS SWIPT UL transmit power for D_n. $E'_n = T_s \zeta |h_n|^2 P_T$ is the harvested energy, $0 \leq \zeta \leq 1$ is the EH efficiency, and P_T is the DL eNodeB power budget.

$$\eta_n = \left(1 - T_s\right) B \log_2 \left(1 + \frac{P'_n \left|h_n\right|^2}{N_0 B + \sum_{i=n+1}^{d} P'_i \left|h_i\right|^2}\right) \tag{4.5}$$

For optimal EE, T_s must be of sufficient time to satisfy Equation (4.6), where $P'_t = 24$ dBm $= 0.25$ W, is the UL device maximum transmission power in LTE-A [117]. It is assumed that $\zeta = 0.5$, as in the DL.

$$P'_n = \left(\frac{T_s}{1 - T_s}\right) \zeta \left|h_n\right|^2 P_T \leq \left(1 - T_s\right) P'_t \tag{4.6}$$

The eNodeB broadcasts the same EH signal to all devices during T_s, where each device harvests energy of P'_n. Since there is a difference in the channel gains (due to Rayleigh fading) for each device, then the separation of each UL signal during the SIC process at the eNodeB is easier because devices located closer to the eNodeB harvest more power.

4.5 NOMA for MMC Networks

The CH of each cluster will be responsible for providing information to the eNodeB of the power requirements within its cluster. Each device (excluding the weakest device) within a cluster will have a transmit power equal to the UL device maximum transmission power of $P'_t = 24$ dBm $= 0.25$ W (mentioned earlier in Section 4.4). The weakest channel device will have power control applied [see Equations (4.9)–(4.10)], if the necessary conditions given in Equations (4.11)–(4.13) are not fulfilled. Power control is used to maintain power level distinctness, but this reduced transmit power cannot be re-allocated to the stronger channel devices to increase individual throughputs, unlike in the DL [2]. The considerations on power allocation leads to a simpler formulation of a convex optimisation problem [76], as shown in Equation (4.7). η_n is given in Equation (4.2).

$$\begin{aligned} &\underset{\alpha_n}{\text{maximise}} && \eta_n \\ &\text{subject to:} && P_n \leq P'_t \\ & && \sum_{n=1}^{d} \alpha_n = 1 \\ & && P_n \geq 0, \quad n = 1, \ldots, d \end{aligned} \tag{4.7}$$

In [76], KKT optimality conditions were used to derive closed-form solutions for optimal power allocation (see Subsection 2.5.7), based on an optimisation problem like Equation (4.7). The large number (i.e. 2^{3m-1} for a cluster of size m) of Lagrange multiplier combinations required to be checked to satisfy the KKT conditions, were reduced to a general three power allocation equations, since $P_n \geq 0$. Each set of power allocation equations has a corresponding set of necessary conditions that must be satisfied, as mentioned in Subsection 2.5.7. The power allocation equations and corresponding necessary conditions consist of $\phi_i = 2^{\beta} - 1$, where $\beta = (R_i \;/\; \omega B_{\text{RB}})$. R_i is the minimum rate requirement, ω is the number of RBs allocated to a cluster, and $B_{\text{RB}} = 180$ kHz is the bandwidth of a RB. ω is equal to the size of the particular cluster, to facilitate a fair comparison between NOMA and OMA. ϕ_i influences which set of power allocation equations is selected for each cluster. Therefore, ϕ_i also influences Equations (4.2) and (3.9), because of R_i. The UL power budget for each cluster is $\omega P_t'$.

As an example, there are two possible power allocations that can be used for a 2-device NOMA cluster, as shown in Equations (4.8)–(4.10). Note that $\gamma_n = \sqrt{|h_n|^2} \;/\; N_0 B_{\text{RB}}$, is the normalised channel gain [76].

$$P_1 = P_2 = P_t' \tag{4.8}$$

$$P_1 = P_t', \quad P_2 = \frac{P_t' \gamma_1}{\phi_1 \gamma_2} - \frac{\omega}{\gamma_2} \tag{4.9}$$

$$P_1 = P_t', \quad P_2 = \frac{P_t' \gamma_1}{\gamma_2} - \frac{P_{\text{tol}}}{\gamma_2} \tag{4.10}$$

The corresponding necessary conditions for Equations (4.8)–(4.10), are shown in Equations (4.11)–(4.13) respectively for $m = 2$ and $n = 1, 2$. For example, $(\mathbf{C_1^2})$ is the first necessary condition for a 2-device cluster, that is checked in Equations (4.12) and (4.13). Only one set of power allocations can be used at one time, depending on the value of R_i. If R_i is too high, then none of the necessary conditions are satisfied, meaning that R_i cannot be met for all cluster devices.

$$(\mathbf{C_1^2}) : P_n \gamma_n - \sum_{j=n+1}^{d} \phi_n P_j \gamma_j - \phi_n \omega > 0, \; n = 1, 2 \tag{4.11a}$$

$$(\mathbf{C_2^2}) : P_1 \gamma_1 - P_2 \gamma_2 - P_{\text{tol}} > 0 \tag{4.11b}$$

$$(\mathbf{C_1^2})\ n = 2,\ (\mathbf{C_2^2}),\ \textbf{AND}\ P_2 < P_t' \tag{4.12}$$

$$(\mathbf{C_1^2})\ \textbf{AND}\ P_2 < P_t' \tag{4.13}$$

The steps for the proposed NOMA algorithm for MMC clusters are given in **Algorithm 4.2**. Note that SWIPT in step 4 of **Algorithm 4.2** is an extension of the proposed **Algorithm 4.1** and the optimal power allocation in step 3. Thus, the resulting solution from the optimisation problem of Equation (4.7)[2] is not applicable to SWIPT.

Algorithm 4.2: NOMA for MMC Network

1. Calculate h_n for each device [as in Equation (2.9)], based on their distance d_n from the eNodeB.
2. Within each cluster, sort the devices in descending $|h_n|^2$, so that $\omega P_t'$ can be allocated according to [76, Table 2].
3. Determine which set of power allocation equations to use for a cluster, based on size and the corresponding set of satisfied necessary conditions.
4. Calculate the required T_s to harvest enough energy for each device, to satisfy Equation (4.6).

4.6 Results and Discussion

An example network of 50 M2M devices is shown in Figure 4.3, as used in the DL.

4.6.1 K-means Clustering

The average silhouette value for $k = \{13, 14, \ldots, 25\}$ (see Step 1 in **Algorithm 3.2**) was first computed, with the highest being the optimal k value. The k strongest channel gain

[2]Equation (4.7) can easily be extended to include SWIPT (i.e. results shown in Figures 4.10, 4.12–4.13a), but the enhanced K-means clustering algorithm developed in **Algorithm 4.1** (i.e. results shown in Figures 4.14–4.15), is the main outcome of this thesis.

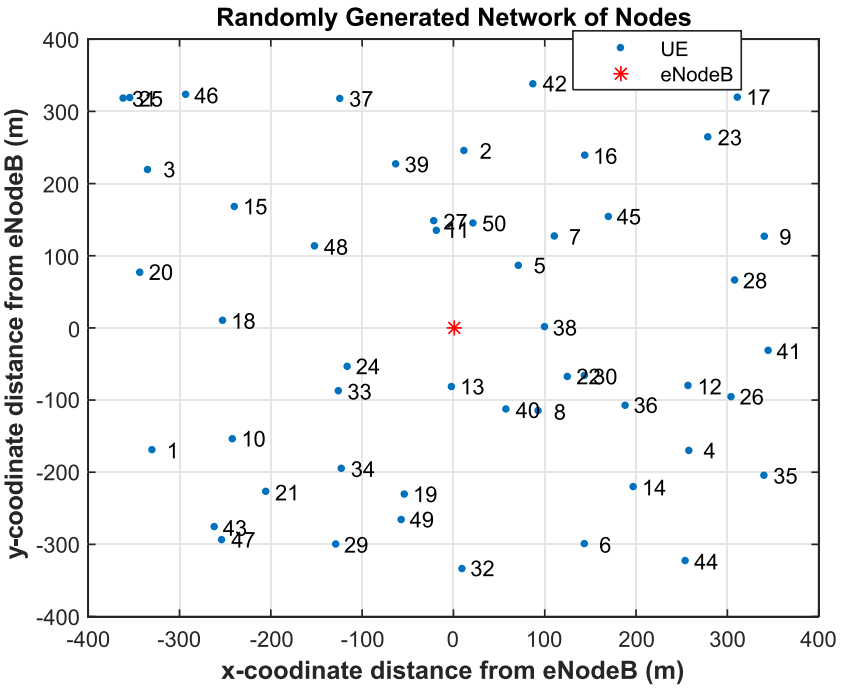


Figure 4.3: An example network of 50 devices, labelled from D_1 to D_{50}.

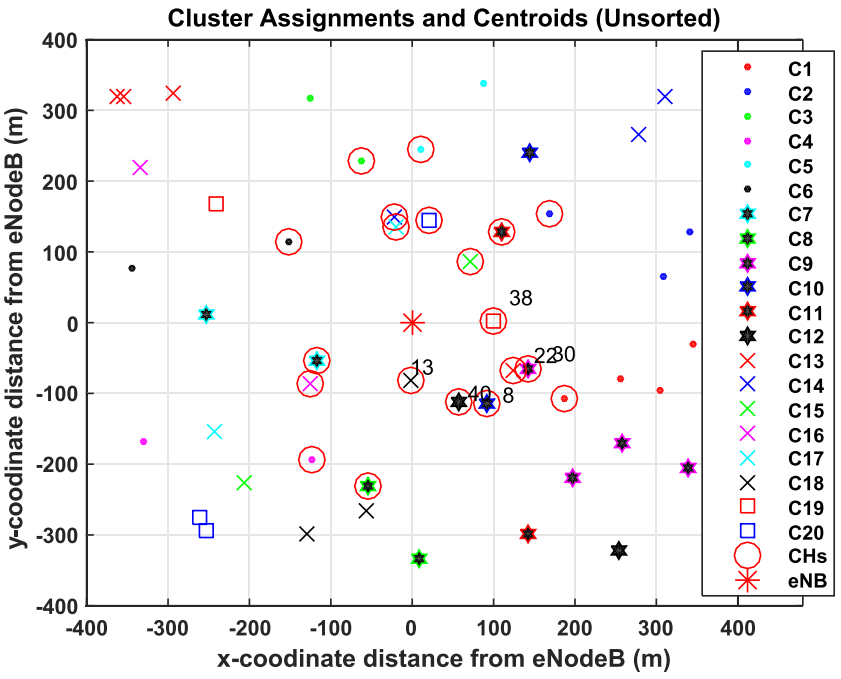


Figure 4.4: The example network from Figure 3.4, split into $k = 20$ clusters by using our enhanced K-means algorithm.

devices were excluded from the network in Figure 4.3. The K-means algorithm was run 25 times for the remaining network with $k = 20$ (as shown in Figure 4.4), with differing initial centroid locations, for a higher probability of finding the best clustering solution. Each of the excluded k strongest channel gain devices were assigned to the appropriate cluster to enhance the network sum throughput, in accordance with 'User Pairing'.

However, it was found during the simulations that 4-device clusters were likely to experience outage if there was a small difference in channel gains between the 2 weakest devices and the second strongest device within a particular cluster. This can be explained by considering ($\mathbf{C_2^m}$) in Equation (2.17), where the normalised channel gains (i.e. γ_n) are more dominant compared to the power allocated to each device (i.e. P_n). This was not an issue during DL simulations in [2], because only the normalised channel gain of the immediate stronger device of the target device (i.e. γ_{n-1}) is considered, as shown in Equation (2.16). This issue was solved with the additional step in **Algorithm 4.1**, where the revised cluster result is shown in Figure 4.5. Referring to Figure 4.4, the second strongest device in the 4-device clusters of C1, C9 and C13, were re-assigned to the 2-device clusters of C12, C11 and C16, respectively.

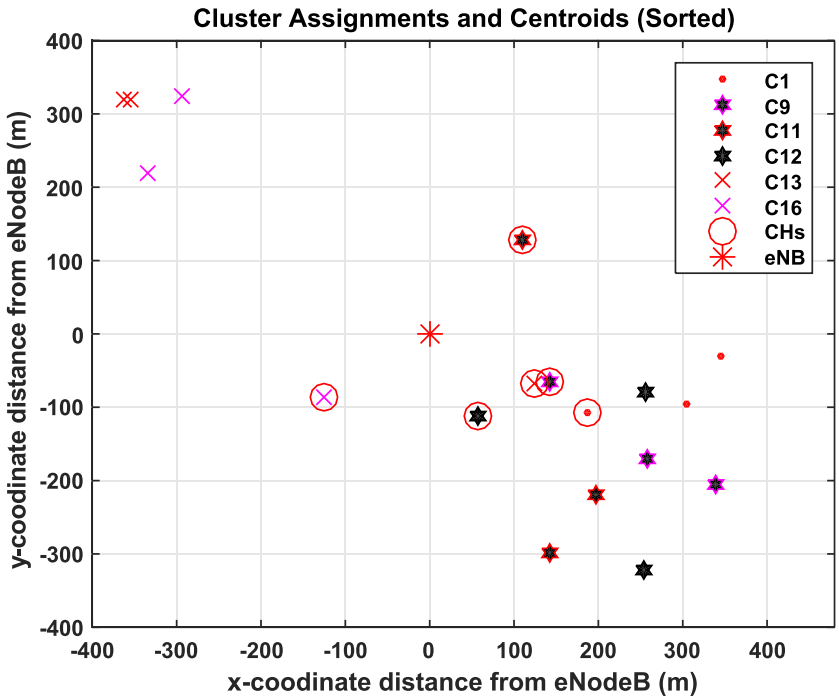


Figure 4.5: The new cluster result after applying the additional step in **Algorithm 4.1**. Only the revised clusters are shown for ease of comparison with Figure 4.4.

Table 4.1: MATLAB UL Simulation Parameters

Parameters	Setup
Carrier Bandwidth, W	10 MHz
Number of available RBs	50
Transmit Antennas	1
Receive Antennas	1
DL Transmit Power Budget, P_T	46 dBm [117]
Maximum UL Transmit Power/device, P'_t	24 dBm [117]
Minimum SIC threshold, P_{tol}	10 dBm [76]
Path loss exponent, α	3.76 [117]
ω	2–4 [76]
Minimum rate requirement, R_i	100 kbps
Energy harvesting efficiency, ζ	0.5 [58]

4.6.2 NOMA

The simulation parameters shown in **Table 4.1**[3] were used for both NOMA and OMA. X from Equation (4.14) was generated (see page 36 for more explanation), and then $|h_n|$ was computed for every device within the network. For each device, $P'_t = 24$ dBm $= 0.25$ W, while $P_T = 46$ dBm $= 39.8107$ W, for an eNodeB in LTE-A [117].

$$h_n = X\sqrt{\left(\frac{1}{2(d_n)^{\alpha}}\right)} \tag{4.14}$$

Using Equation (4.4), network sum throughput was determined while utilising NOMA and OMA respectively, by performing the summation of Equations (4.2) and (4.3). The resulting sum throughput for each cluster, is shown in Figure 4.6. It can be observed that the sum throughput of each NOMA cluster is higher than the corresponding OMA cluster.

[3]Note that the simulation uses the normalised channel gain (i.e. $\gamma_n = \sqrt{|h_n|^2} \;/\; N_0 B_{\text{RB}}$) instead of N_0, based on the simulation results in [76].

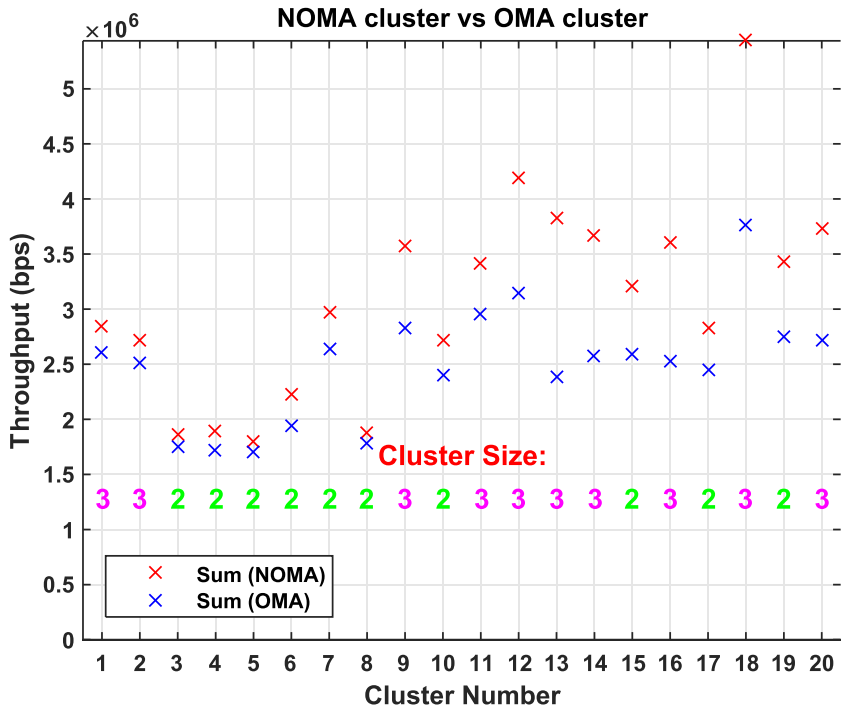


Figure 4.6: Sum throughput comparison between NOMA and OMA clusters formed using the enhanced K-means algorithm, for $R_i = 100$ kbps.

4.6.3 Enhanced K-means vs Traditional K-means

The NOMA network sum throughput for the enhanced K-means was compared with the traditional K-means and Hierarchical Clustering as shown in Figure 4.7. The simulation was run 25 times for both schemes and Hierarchical Clustering, and the results were averaged out to mitigate the differing of the 'best clustering solution' from each simulation run. The averaged results are summarised as follows:

- Hierarchical was around 0.17% higher than OMA. The traditional K-means was around 0.57% higher than OMA.
- The enhanced K-means NOMA network sum throughput was around 24.77% higher than OMA. This represents an improvement of around 24.06% (on average) over the traditional K-means and 24.49% (on average) over Hierarchical, when using the proposed enhanced K-means.

The proposed scheme achieves a higher percentage gain over OMA compared to traditional K-means, because 'User Pairing' was used to enhance the network sum throughput and traditional K-means tends to make some clusters of size 4 or larger.

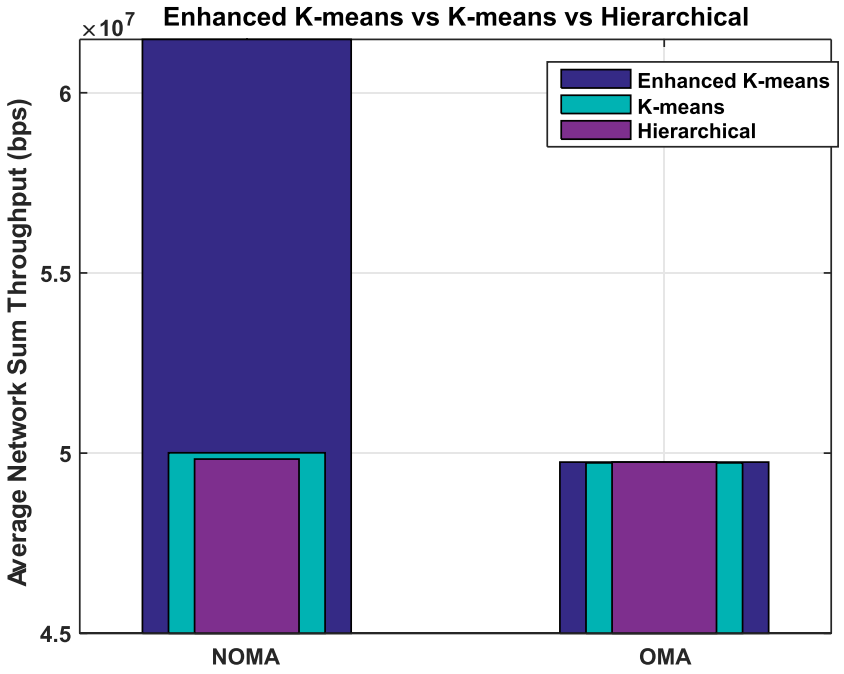


Figure 4.7: NOMA vs OMA network sum throughput (on average) comparison between the enhanced K-means, traditional K-means and Hierarchical, for $R_i = 100$ kbps.

Referring to Figure 4.3, the traditional K-means would most likely assign D_8, D_{13}, D_{22}, D_{30}, D_{38} and D_{40} into one cluster. However, each of these devices are one of the k strongest channel devices, which means they would instead be assigned as CHs to separate clusters (as shown and labelled accordingly in Figure 4.4) when the proposed scheme is utilised. As mentioned earlier, 4-device clusters were likely to experience outage if there was a small difference in channel gains between the 2 weakest devices and the second strongest device within a particular cluster. The likelihood of outage with the traditional K-means and Hierarchical was reduced, by increasing the number of clusters formed. However, this resulted in many un-clustered devices, which led to minimal throughput advantage over OMA.

As another example, the proposed scheme was run for the Figure 4.8, with a different distribution of device locations in relation to the eNodeB. The averaged results are summarised below:

- The enhanced K-means NOMA network sum throughput was 22.78% higher than OMA, where the network sum throughput gain is 1.50% lower (essentially similar) than the network in Figure 4.3.

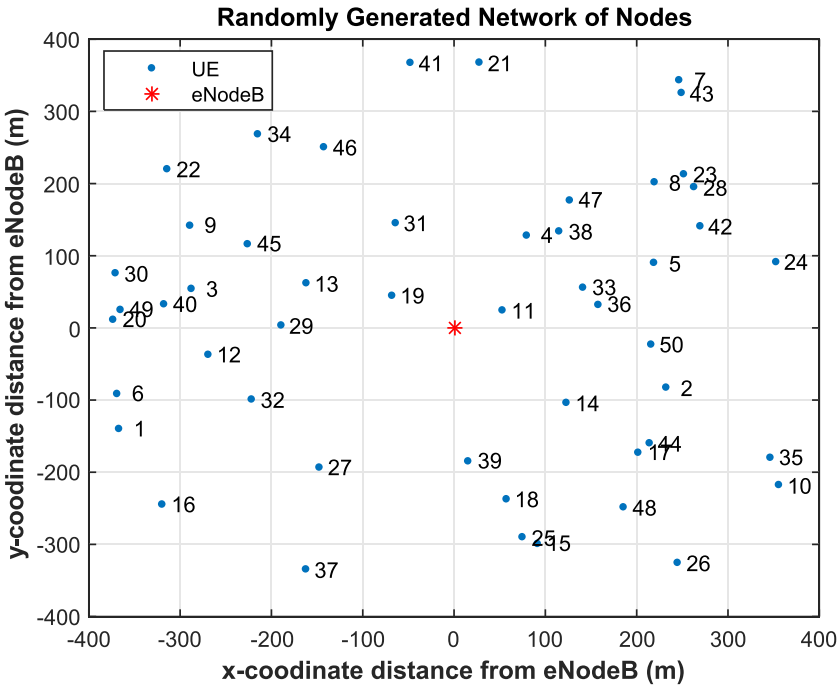


Figure 4.8: An example 2nd network of 50 devices, labelled from D_1 to D_{50}.

- The OMA network sum throughput is 5.00% lower than the network in Figure 4.3, indicating a lower average channel gain for devices.

4.6.4 Varying Minimum Rate Requirements

As mentioned in Section 4.5, only one set of power allocation equations can be used at one time. The selected set of power allocation equations is based on the value of R_i and also how different the $|h_n|^2$ of devices are within a particular cluster.

4.6.4.1 2-Device and 3-Device Cluster Example with SWIPT

A 2-device cluster with $|h_1|^2 = 40$ dB and $|h_2|^2 = 30$ dB was considered, where there was a small enough difference in channel gains, to easily observe the effect of power control at the weaker device. This contrasts with DL NOMA, where a large difference was chosen in [2]. $R_i = \{100, 120, \ldots, 1500\}$ kbps was set as shown in Figure 4.9. It can be observed:

- The NOMA sum throughput is constant for 100 kbps $\leq R_i \leq$ 1250 kbps, because the power allocated to both devices are similar according to Equation (4.8).

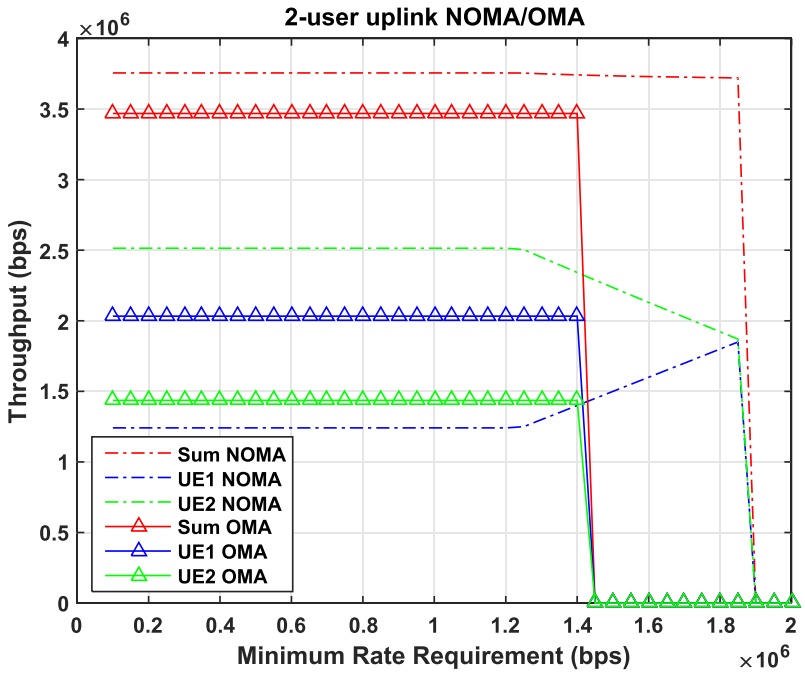


Figure 4.9: NOMA vs OMA at varying R_i with $|h_1|^2 = 40$ dB and $|h_2|^2 = 30$ dB. **NOTE: the throughput is set to 0 bps when the system is in 'outage'.**

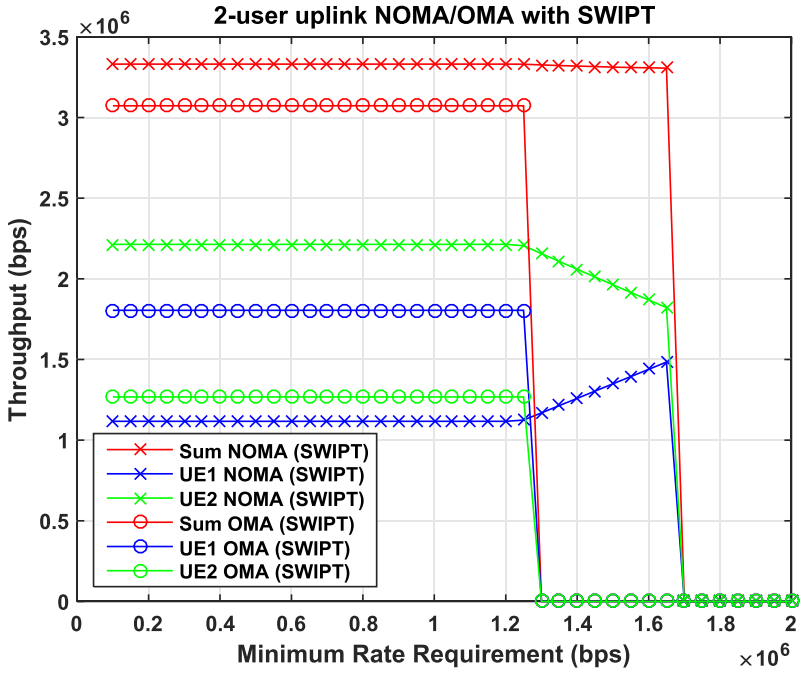


Figure 4.10: The same parameters are used as in Figure 4.9, but with SWIPT included. **NOTE: the throughput is set to 0 bps when UL transmission is not feasible.**

- The sum throughput starts to decrease when $R_i > 1250$ kbps, since Equation (4.9) or Equation (4.10) is used instead to apply power control.
- From $R_i > 1850$ kbps for NOMA, and $R_i > 1400$ kbps for OMA, Equation (4.8)–(4.10) can no longer be satisfied as R_i cannot be met for both devices, as in the system is in 'outage'.

Using Equation (4.6) with $T_s = 0.1$, Figure 4.10 shows that:

- UL transmission is feasible for 100 kbps $\leq R_i <$ 1700 kbps, if it is assumed that each device's battery is finite.
- There is a slight decrease in the NOMA sum throughput (comparing to Figure 4.9) due to the normalised transmission time period being split between EH and UL transmission.

Extending to a 3-device cluster with $|h_1|^2 = 40$ dB, $|h_2|^2 = 30$ dB, and $|h_3|^2 = 16$ dB, Figure 4.11 shows a similar throughput trend as for 2-device clusters, but the channel gain difference was too large to see the effect of power control at the weakest device. With

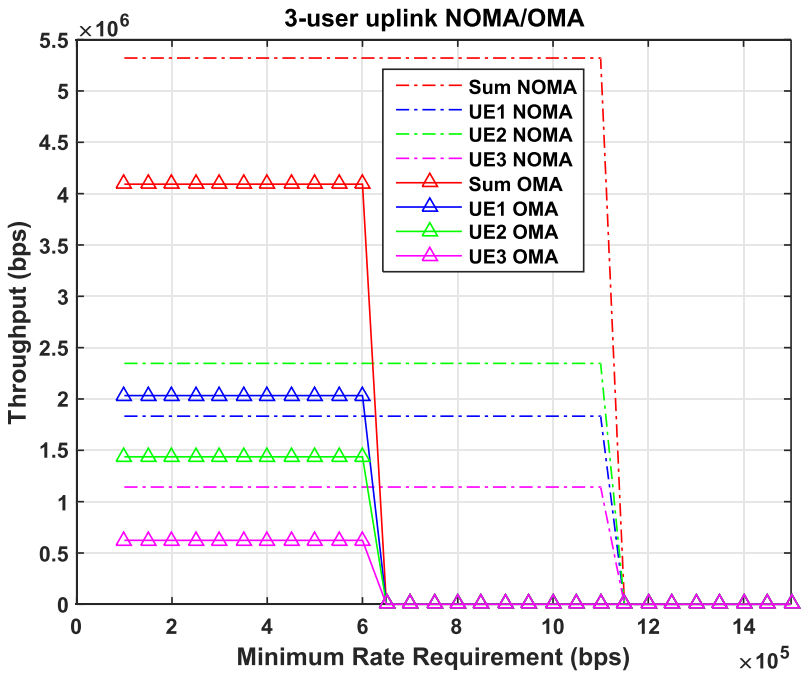


Figure 4.11: NOMA vs OMA at varying R_i with $|h_1|^2 = 40$ dB, $|h_2|^2 = 30$ dB and $|h_3|^2 = 16$ dB.

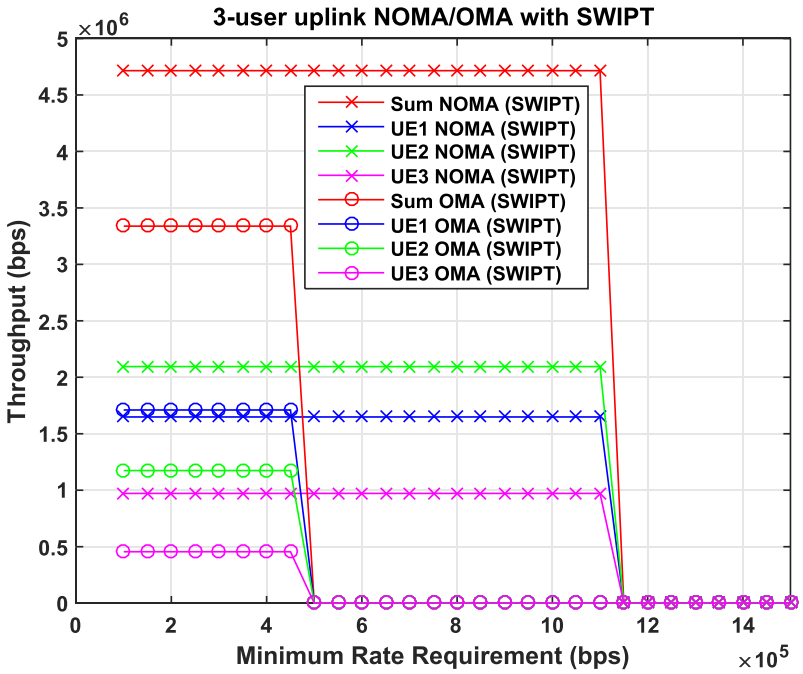


Figure 4.12: The same parameters are used as in Figure 4.11, but with SWIPT included.

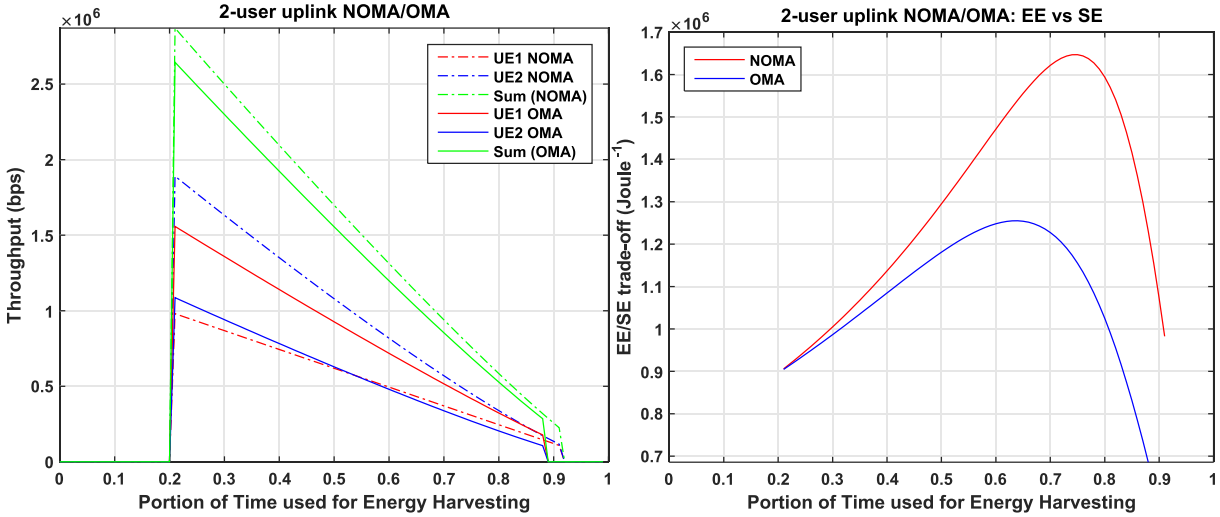
$T_s = 0.1$, Figure 4.12 shows that UL transmission is feasible for 100 kbps $\leq R_i <$ 1150 kbps, assuming that each device's battery is finite. **This is a promising result for utilising UL transmission for cluster sizes of larger than 2, as to the best of our knowledge, other papers dealing with SWIPT (without using MTCG) have only considered clusters of size 2.**

4.6.4.2 Energy-Efficiency (EE)

In Figure 4.13a, we set $T_s = \{0, 0.01, \ldots, 0.99\}$ for a 2-device cluster, with $|h_1|^2 = 40$ dB and $|h_2|^2 = 30$ dB. At $R_i = 100$ kbps, UL transmission is feasible for $0.2 < T_s < 0.92$, where NOMA is superior to OMA within that region. **EE** is related to the spectral efficiency (**SE**) as shown in Equation (4.15), where P_{static} is the circuitry power consumption [70], and P_T is defined as in Subsection 4.6.2.

$$\textbf{EE} = \frac{\eta_T}{P_T + P_{\text{static}}} = \textbf{SE} \times \frac{B}{P_T + P_{\text{static}}} \tag{4.15}$$

Therefore, there is a trade-off between **EE** and **SE** (as shown in Figure 4.13b), where increasing the **SE** past the optimal **EE** point, will lead to a decrease in the **EE** [70]. As T_s is increased towards 0.92, the throughput degrades but more energy is harvested.



(a)

(b) EE/SE trade-off based on Figure 4.13a parameters.

Figure 4.13: NOMA vs OMA at varying T_s for SWIPT, for $R_i = 100$ kbps.

Excess harvested energy can be stored for later use, in a supercapacitor or a short-term efficiency battery [57].

Using SWIPT can even reduce the *carbon footprint*, if an EH-Base Station (EH-BS) provides harvested 'green' energy. Green energy is an important environmental consideration for the IoT and cellular networks in general, but there are current limiting monetary factors to consider, such as the purchase quantity and price of green energy [56].

4.6.4.3 Enhanced K-means

The enhanced K-means algorithm was run at $R_i = \{100, 120, \ldots, 1500\}$ kbps, as shown in Figure 4.14, based on the network in Figure 4.3. It can be observed that our enhanced K-means network sum throughput is higher than OMA for all R_i. The network sum throughput starts to decrease as expected when a new set of power allocation equations are used. However, it is obvious that the necessary conditions for 3 and 2-device clusters, respectively, can no longer be satisfied as R_i is increased towards 1500 kbps.

4.6.4.4 Different Sized Clusters

The 3-device clusters in Figure 4.3 begin to experience outage from $R_i > 420$ kbps, while the 2-device clusters from $R_i > 780$ kbps. Therefore, it is easier to satisfy the necessary

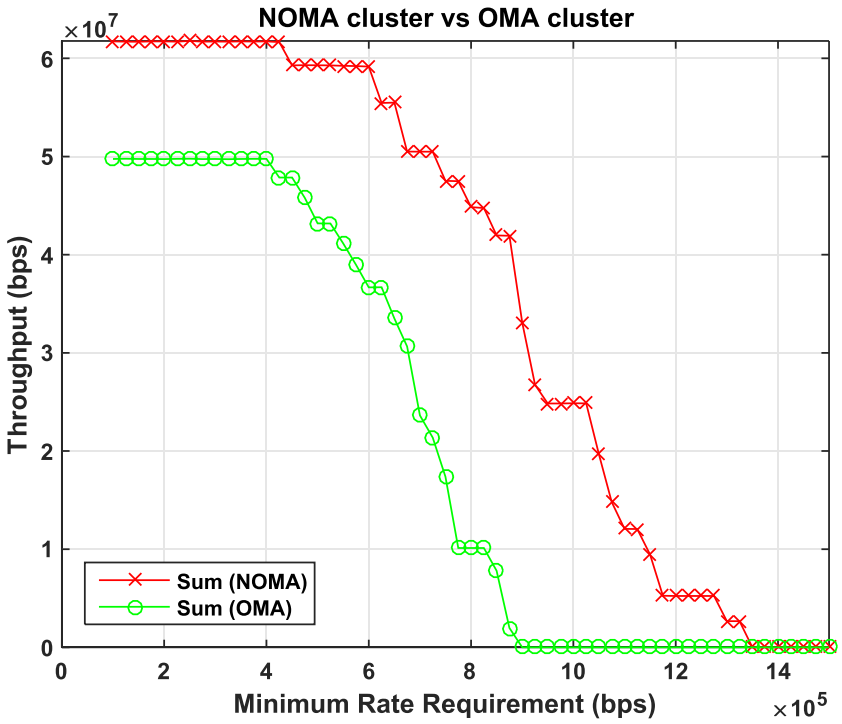
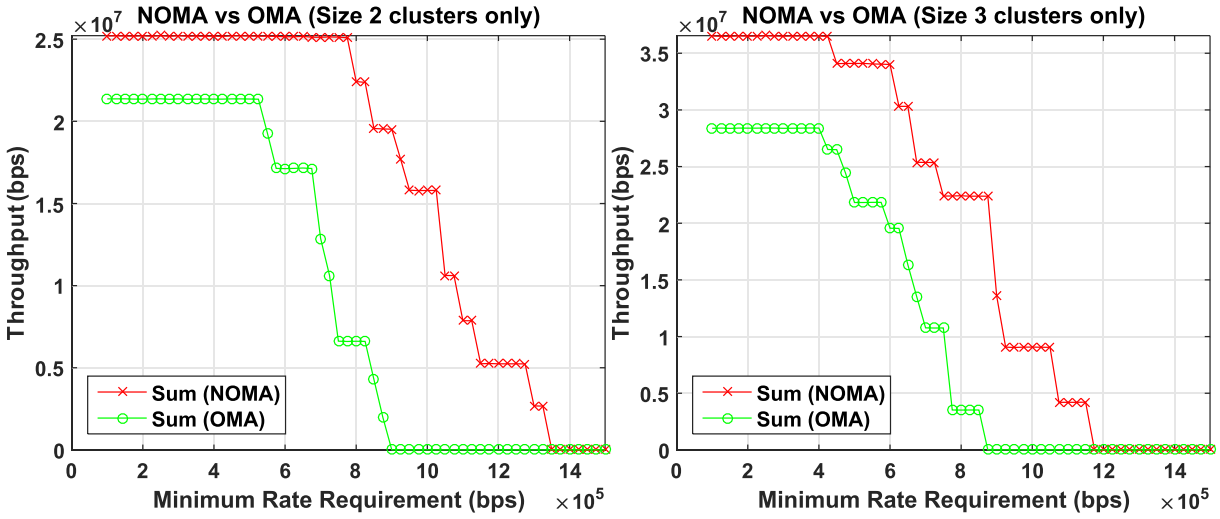


Figure 4.14: NOMA vs OMA network sum throughput with varying R_i. **NOTE: the steep drops in sum throughput is due to a particular cluster or clusters being in 'outage'.**

conditions for smaller cluster sizes, which means that larger networks are limited to lower values of R_i. This can be seen in Figures 4.15a and 4.15b for 2 and 3-device clusters, respectively, from Figure 4.3. Refer to Figure 4.6 for the cluster sizes for the network in Figure 4.3.

This R_i limitation would not be such a problem, as these MMC devices are expected to be transferring small amounts of data. It is much more important to be able to simultaneously accommodate the billions of devices within an MMC network. The enhanced K-means algorithm achieves this as it avoids the possibility of an overload and access problem by placing devices into clusters and utilising NOMA. However, when $R_i > 420$ kbps, some of the clusters will be in outage for the Figure 4.3 network. Having smaller cluster sizes would increase the R_i for when the system is in outage for NOMA, and there is a particular cluster size (which is 2–3 based on our results) where NOMA is superior to OMA at all network conditions.



(a) Size 2 clusters only **(b)** Size 3 clusters only

Figure 4.15: Figure 4.14 but with 2 and 3-device clusters only.

4.7 Combined DL/UL NOMA with SWIPT

Based on **Algorithm 3.2** and **Algorithm 4.1** for the DL and UL solutions, respectively, in Chapters 3 and 4, it is clear that the UL clustering solution must be used for a combined DL/UL solution, as there are some throughput problems observed in the UL. This will result in the same clustering outcome for the combined DL/UL solution, with a reduced complexity, as the cluster memberships do not need to be modified for every time DL NOMA and UL NOMA is required.

It is proposed that the normalised transmission time slot is split into two parts, where the first part consists of DL NOMA (as in **Algorithm 3.2**) and accompanied by PS SWIPT, and the second part consists of UL NOMA (as in **Algorithm 4.1**) and accompanied by TS SWIPT. To simplify the EH process (which can also be used later in the following Chapter 5), it is assumed that an energy-rich (ER) node (which is equipped with a stable power supply) is available for each cluster [61]. The ER node will broadcast an EH signal to the cluster members, so that UL transmission is feasible for the devices (assuming that each device's battery is finite). MMC devices will mostly communicate in the UL, while URLLC devices will mostly communicate in the DL, where the IoT consists of a combination of MMC and URLLC devices.

4.8 Conclusion

In this chapter, an enhanced K-means algorithm with UL NOMA and SWIPT was proposed to fulfil the requirements of MMC devices (such as sensor nodes), which have the QoS requirements of massive connectivity, small packets, low data rate, high EE, low energy consumption, etc. However, there were some throughput problems observed in the UL, which resulted when the difference between the 2 weakest channel devices and the second strongest device within a 4-device cluster was too small.

The proposed scheme used the same procedure as in the DL, but had an additional step as given in **Algorithm 4.1**. This additional step involved re-assigning the second strongest device in every 4-device cluster to the closest 2-device cluster.

The earlier mentioned clustering was also used to reduce the power consumption of the NOMA SIC process compared to an un-clustered MMC network, although this is not as much of a concern compared to the DL, since SIC is carried out at the eNodeB. In order to address the EE of UL transmission, the TS SWIPT EH technique was used to mitigate the power consumption of UL transmission.

The proposed scheme was shown in Figure 4.6 to have a higher NOMA cluster sum throughput for the $k = 20$ clusters formed, at a minimum rate requirement of $R_i = 100$ kHz. It was also shown in Figure 4.7 to have a 24.77% higher average NOMA network sum throughput (over 25 simulation runs) over OMA, a 24.06% improvement over traditional K-means and a 24.49% improvement over Hierarchical Clustering, at a minimum rate requirement of 100 kHz. For 2-device clusters, it was shown in Figure 4.10, that UL transmission is feasible (assuming that each device's battery is finite) for 100 kbps $\leq R_i <$ 1700 kbps, with a slight decrease in the NOMA sum throughput. This was when the portion of the normalised transmission time period, $T_s = 0.1$. It was also shown in Figure 4.13a, that UL transmission is feasible for $0.2 < T_s < 0.92$, with $|h_1|^2 = 40$ dB and $|h_2|^2 = 30$ dB, at $R_i = 100$ kHz. Extending to 3-device clusters, it was shown in Figure 4.12, that UL transmission is feasible for 100 kbps $\leq R_i <$ 1150 kbps. The proposed scheme was shown in Figures 4.14 and 4.15 to have a higher network sum throughput improvement over OMA compared to the traditional K-means with 100 kbps $\leq R_i \leq$ 1500 kbps. Applying TS SWIPT showed promising results in increasing the EE by mitigating the power consumption of SIC and of UL transmission.

Chapter 5

Enhanced NOMA HARQ Scheme

5.1 Introduction

5.1.1 Motivations

As mentioned in Section 1.3, this thesis will also focus on enabling the URC operating mode, which has the QoS requirement of ultra-reliability. The ultra-reliable probability of $P_{\mathrm{UR}} \geq 99.999\%$ can be easily achieved if the given deadline is sufficiently long [23]. However, the URLLC operating mode has an additional QoS requirement of ultra-low latency (i.e. ≤ 1 ms), which makes it more challenging to meet the ultra-reliable probability requirement. Severe consequences such as high financial cost, injury or even death, could result from non-reliable data transmission for URLLC applications, such as mission critical industrial control, medical, and V2X. Some of the potential solutions mentioned in Subsection 2.7, are the use of the 5G longer subcarrier spacing, mini-slots (i.e. 2, 4 or 7 OFDM symbols per TTI) and grant-free HARQ (by utilising slot repetition in 5G).

5.1.2 Solution

In this chapter, an enhanced NOMA scheme with HARQ is proposed, to fulfil the requirements of the URLLC applications mentioned in Subsection 5.1.1, which are used as motivations. The same clustering procedure in **Algorithm 4.1** was followed, to resolve the throughput problems observed in the UL. Thus, this chapter is an extension from the enhanced K-means clustering result observed in Chapter 4. The proposed algorithm involves:

1. designating the network of M2M devices devices into k clusters, using the enhanced K-means clustering scheme in **Algorithm 4.1**.

2. the procedure as described later in Section 5.4 is carried out with DL NOMA and PS SWIPT, then the ER node broadcasts an EH signal towards all devices within a cluster, and lastly UL NOMA.

3. the eNodeB performs RTX with one of the RV of the superimposed signal, and the strong channel gain device performs cooperative relaying NOMA after successfully decoding the weak channel gain device's signal.

This chapter is organised as follows: the fundamentals for HARQ and cooperative relaying NOMA, and the NOMA HARQ state of the art, are given in Section 5.2. Section 5.3 establishes the system model and states the assumptions. The proposed enhanced NOMA HARQ scheme is presented in Section 5.4. The outage probability, SINR and achievable rate equations, used in Section 5.6 are shown in Section 5.5. A comparison between the proposed scheme and the LTE-A HARQ scheme based on the MATLAB simulation results is shown Section 5.6. Section 5.7 concludes this chapter.

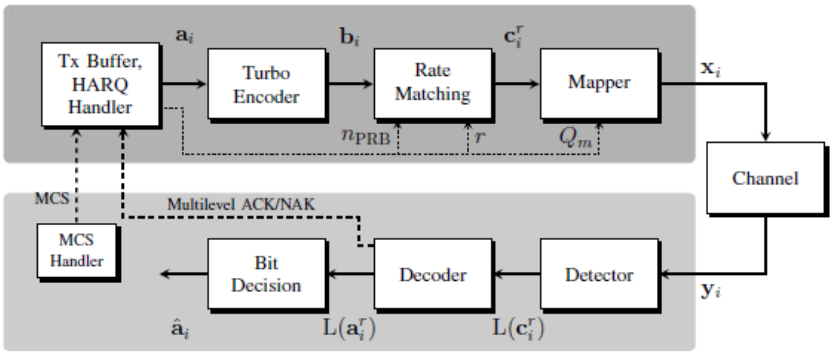


Figure 5.1: LTE-A HARQ process utilising a turbo encoder and multi-level ACK/NAK. © 2014 IEEE, from [118].

5.2 HARQ and Cooperative Relaying NOMA

5.2.1 HARQ LTE Process

HARQ[1] is a Stop and Wait (SAW) process in LTE-A [6], where it requires feedback before it can continue. A TB is constructed as shown in Figure 5.1 (on page 96), by passing a binary data sequence of length N_a through a turbo encoder (with a mother code rate of $N_a/N_b = 1/3$), resulting in a codeword of length N_b. The encoded codeword is passed to a rate matcher, which outputs a codeword of length N_c, depending on the RV number and the number of available physical RBs. The rate matched codeword is passed to an OFDM bit mapper, to 2-bit Quadrature PSK (QPSK), and 4-bit/6-bit Quadrature Amplitude Modulation (16QAM/64QAM). The Tx then sends the TB to the Rx over a channel. The 8 ms LTE-A HARQ process is shown in Figure 5.2 [3] and in **Algorithm 5.1**.

Algorithm 5.1: LTE-A HARQ Process

1. The received TB is decoded by the Rx, by using a Cyclic Redundancy Check (CRC) [119]. If the CRC is successful, Acknowledgement (ACK) feedback is encoded. Otherwise, Negative ACK (NAK) feedback is encoded.

2. The ACK/NAK feedback is transmitted to the Tx.

3. The received feedback is decoded by the Tx. If decoding is successful, the next TB is constructed. Otherwise, the RTX TB is constructed. These TBs are then encoded for transmission to the Rx.

4. The next TB or RTX TB is transmitted to the Rx.

One problem with using a single HARQ SAW process is that the process has to wait for feedback during the majority of the 8 ms round trip, which can be observed in Figure 5.2. This is why LTE-A uses 8 parallel HARQ processes in the uplink direction, or up to 8 in the downlink direction [6].

[1] parts of Subsection 5.2.1 was previously published in **Section 2.3** of my MRes thesis in 2015 [4].

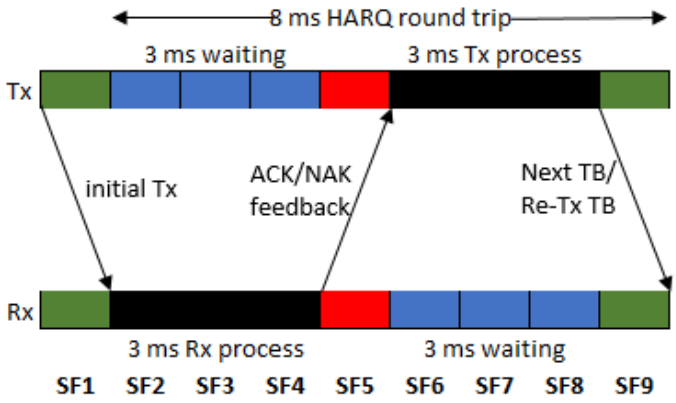


Figure 5.2: The 8 ms LTE-A HARQ process. Data Tx/RTX in green (SF1/SF9), Tx/Rx processing in black (SF6–8/SF2–4), ACK/NAK feedback in red (SF5), and the Tx/Rx feedback waiting period in blue (SF2–4/SF6–8). © 2017 IEEE, from [3].

5.2.2 Adaptive HARQ (A-HARQ) for URC in 5G

HARQ is sufficient for most applications, and can enable the URC operating mode, provided the deadline is sufficiently long. However, it is likely that the system will be in 'outage' if the channel conditions are poor enough, since the limit of 4 HARQ RTXs will be reached and thereby lead to excessive delay [3].

Applications in low SNR environments, that also require URLLC have been addressed in [3][2], where an Adaptive HARQ (A-HARQ) scheme was proposed by using better quality sub-bands for RTX, with resources dynamically allocated using Channel Quality Indicator (CQI) reports. A-HARQ is based on Type-II HARQ to fulfil URLLC applications. The current LTE-A HARQ scheme is referred to as 'legacy HARQ'.

To ensure a quicker RTX compared to the legacy HARQ, A-HARQ involves pre-constructing and pre-encoding differing RVs with different combinations of systematic bits and parity bits, and then storing them in the Tx buffer. The existing TTI bundling from LTE [6] was utilised to increase the number of RTX within a time period of 4 ms, since the goal was to reduce the delay incurred by multiple RTX when using the legacy HARQ. CQI reports from LTE-A [6] are also utilised to assist the eNodeB in selecting the best sub-bands to send the RTX, as the use of TTI bundling for RTX could still fail if the same poor quality channel is used. The proposed A-HARQ RTX process (as shown in Figure 5.3) is as follows:

[2]parts of Subsection 5.2.2 was previously published in my VTC2017-Spring paper [3].

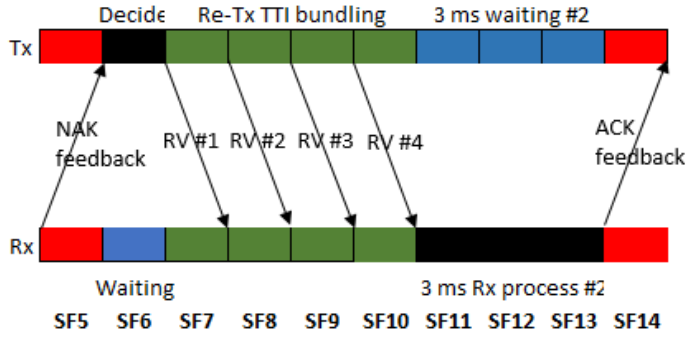


Figure 5.3: Cabrera *et al.*'s proposed A-HARQ scheme. The Tx decodes the ACK/NAK feedback for 1 ms (SF6), then sends RV #1 to RV #4 (SF7–10) in response to the NAK. After the 4 RVs are received, the Rx process #2 runs (SF11–13), which includes the decoding of each received RV, CRC, and the encoding of ACK feedback. © 2017 IEEE, from [3].

- The eNodeB transmits RV #1 to RV #4, in consecutive subframes SF7–SF10, using TTI bundling.
- The Rx receives RV #1 to RV #4 in consecutive subframes SF8–SF11 and decodes each RV in consecutive subframes SF9–SF12, and then stores the received TBs in the Rx buffer.
- After all the RTX TBs in a TTI bundle are stored in the Rx buffer, the Rx applies packet combining on the previously received erroneous TB and RV #1 to RV #4 in SF13 to attempt to decode the packet.

The probability of a successful decoding should be very high since there are many different combinations of systematic and parity bits being combined together.

One advantage with using TTI bundling, is that only 1 set of control signalling is required. This reduces the control signalling overhead required for resource allocation and HARQ ACK/NAK feedback [3]. The results show that A-HARQ incurs around 35% less delay than the legacy HARQ, with a slight decrease in throughput, for the low SNR region of 0–2 dB. A-HARQ is therefore very useful for applications within very harsh signal environments.

5.2.3 HARQ NOMA Schemes

HARQ schemes in NOMA have been investigated since 2008 in [120], but have renewed interest due to NOMA being the strong candidate MA technique for 5G cellular networks [79, 121–126]. However, HARQ has been shown to be more complicated in NOMA compared to OMA, because of the interference experienced from other users.

5.2.3.1 Incremental Redundancy (IR) HARQ

The effect on the outage probability was investigated in [123], when considering the user interference in NOMA, while utilising IR HARQ. IR HARQ is similar to Type-II HARQ, where each RTX has been adapted to the changing channel conditions, but IR HARQ also includes soft combining methods at the Rx. IR was considered because it has been shown to achieve a higher spectral efficiency than CC [127].

Power is normally allocated in NOMA according to the inequalities of $\text{SINR}_2 \geq R_2$ and $\text{SNR}_1 \geq R_1$ for $|h_1|^2$, and $\text{SINR}_2 \geq R_2$ for $|h_2|^2$, where UE1 can decode its own signal after using SIC, and UE2 can decode its own signal by treating UE1's signal as noise. To denote the time slot t, η_n and $|h_n|^2$ are rewritten as η_n^t and $|h_n^t|^2$. The problem is that $|h_n|$ is unknown at the eNodeB, unless the eNodeB receives CSI feedback from the UEs (this is considered in [128]), to be able to do power allocation. HARQ protocols with binary feedback (i.e. using ACK and NAK) is instead considered in [123].

The number of RTXs to decode the signal to user-**A** at user-**B**, $T_{(B),A}$ is given in Equations (5.1) and (5.2) for $|h_1|^2$ and $|h_2|^2$, for the signal to UE2. E.g. $\eta_{1\to 2}^t$ is the rate at UE1 to decode UE2's signal. RTX only occurs if the signal to UE2 cannot be decoded using SIC.

$$T_{(1),2} = \min_T \left\{ T \middle| \sum_{t=1}^{T} \eta_{1\to 2}^t \geq R_2 \right\} \tag{5.1}$$

$$T_{(2),2} = \min_T \left\{ T \middle| \sum_{t=1}^{T} \eta_{2\to 2}^t \geq R_2 \right\} \tag{5.2}$$

The number of RTXs of the signal to UE2 is shown in Equation (5.3), and the number of RTXs of the signal to UE1 is shown in Equation (5.4), for $|h_1|^2$.

$$T_2 = \max\left\{T_{(1),2}, T_{(2),2}\right\} \tag{5.3}$$

$$T_1 = \max\left\{\min_{T}\left\{T\middle|\sum_{t=1}^{T}\eta_{1\to1}^{t} \geq R_1\right\}, T_2\right\} \tag{5.4}$$

The proposed power allocation scheme, termed 'Error Exponent-Power Allocation' (EE-PA), considers statistical CSI of UE1 and $|h_2|^2$ instead, where the RTX power allocation is fixed.

The results in [123] show that their proposed EE-PA scheme can guarantee a certain outage probability for a given total power and a maximum number of RTXs, but it does not maximise the throughput. EE-PA was compared to a simple power allocation scheme (satisfying Equation (5.5), where $P_T = P_1 + P_2$) termed 'equal RX-SINR power allocation' (equal RX-SINR-PA), where the received SNR at user 1 equals the receive-SINR at user 2. It is assumed that $\bar{\alpha} = 1$ and $\bar{\beta} = 1/d^{\alpha}$, where $\alpha = 3$ is the path loss exponent. The proposed EE-PA was shown to have a higher outage probability than equal RX-SINR-PA, but had a higher throughput.

$$\bar{\alpha}P_1 = \frac{\bar{\beta}P_2}{\bar{\beta}P_1 + 1} \tag{5.5}$$

5.2.3.2 Chase Combining (CC) HARQ

The performance of NOMA with HARQ was also investigated in [79], but instead with CC. CC is similar to Type-I HARQ, but CC also includes soft combining methods such as Maximum-Ratio Combining (MRC). Failed decoded signals are stored in a buffer at the Rx and combined with further RTX signals using MRC. The soft combined signal at the Rx is discarded if the maximum number of RTX is reached, as usual for HARQ schemes.

The NOMA outage probabilities for UE1 and UE2 (UE1 is the stronger channel device) after T transmission rounds, are given in Equations (5.6) and (5.7), where $\hat{R}_1$ and $\hat{R}_2$ are the target rates for UE1 and UE2. E.g. $\mathrm{SNR}_{1\to2}^{t}$ is the SNR at UE1 after attempting to decode UE2's signal. The OMA outage probability is shown in Equation (5.8), where $\hat{R}_j$ is the target rate for UE-j.

$$P_{1,T}^{(\mathrm{NOMA})} = 1 - \mathrm{P}\left\{\log_2\left(1 + \sum_{t=1}^{T}\mathrm{SINR}_{1\to2}^{t}\right) > T\hat{R}_2, \right.$$
$$\left. \log_2\left(1 + \sum_{t=1}^{T}\mathrm{SINR}_{1\to1}^{t}\right) > T\hat{R}_1\right\} \tag{5.6}$$

$$P\left(_{2,T}^{\text{NOMA}}\right) = \text{P}\left\{\log_2\left(1+\sum_{t=1}^{T}\text{SINR}_{2\to 2}^{t}\right) \leq T\hat{R}_2\right\} \tag{5.7}$$

$$P\left(_{j,T}^{\text{OMA}}\right) = \text{P}\left\{\frac{1}{2}\log_2\left(1+\sum_{t=1}^{T}\text{SNR}_{j}^{t}\right) \leq T\hat{R}_j\right\} \tag{5.8}$$

The results in [79] show that without using HARQ, the NOMA outage probability for the stronger UE1 is higher than in OMA, and the NOMA outage probability is always lower than in OMA. The outage probability of the stronger UE1 is lower than OMA, when $T \geq 4$ and the UE1 transmission power $= (P_T \times 0.2)$, while the outage probability of the stronger UE1 is lower than OMA, when $T \geq 3$ and the UE1 transmission power $= (P_T \times 0.3)$. The outage probability decreases when T increases, but this increases the delay. For UE1, the transmission power decreases as T increases, when the outage performance of NOMA equals OMA.

However, the proposed scheme in [79] would not be suitable for URLLC applications due to the increasing delay of successive RTX.

5.2.3.3 Truncated ARQ (TARQ)

For URLLC applications, a large number of RTX is not ideal, even if multiple RTX could eventually fulfil the ultra-reliable probability of $P_{\text{UR}} \geq 99.999\%$. A cooperative relaying NOMA Truncated ARQ (TARQ) scheme was proposed in [129] to minimise the delay incurred through RTX, by limiting the number of RTX.

The proposed scheme involves superimposing the signals intended for the relay and for the destination device (i.e. $s_{1,2} = s_1 + s_2$). It is assumed that there is no direct link between the eNodeB and the destination device, due to factors, such as obstacles or there is a large distance between them. The eNodeB transmits $s_{1,2}$ to the relay. If the relay fails to decode s_1 or s_2, then the eNodeB conducts RTX of s_1 or s_2, respectively, with transmission power P_T. The relay forwards the post-SIC s_2 to the destination device. If the destination device fails to decode s_2, then the relay conducts RTX of s_2 with transmission power P_r. The destination device then applies MRC to attempt to decode its own signal. If the RTXs fail, then the eNodeB sends a new $s_{1,2}$ to the relay, rather than continuing to conduct RTX.

The work in [129] was extended in [130], to include two transmission phases of broadcast and cooperative TARQ, with two users U1 and U2 and several relays (R_i) located in

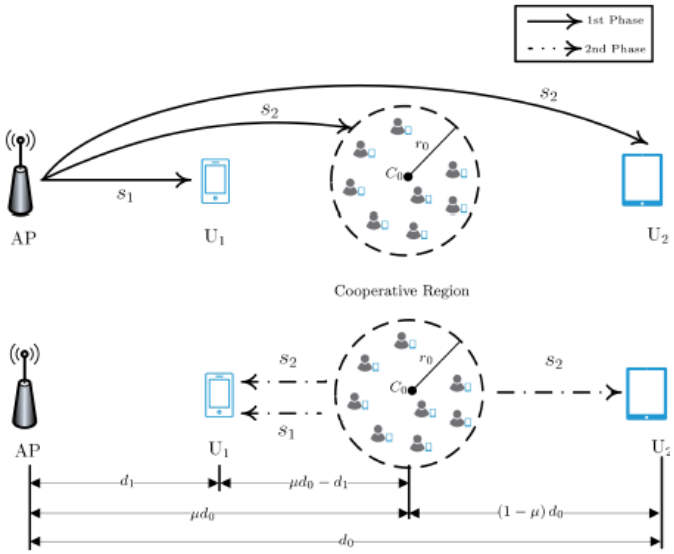


Figure 5.4: Yu's cooperative relaying NOMA TARQ scheme. © 2019 IEEE, from [130].

a disc of radius r_0 (as shown in Figure 5.4).

The eNodeB broadcasts $s_{1,2}$ to U1, U2 and all R_i. The relays that correctly decode s_2 or s_1 are placed into set D_2 or D_1, respectively. If U2 fails to decode s_2, then the best relay (i.e. the relay with the best link to the device that requires RTX, which is UE2 in this case) conducts RTX of s_2. If U1 fails to decode s_2, then the best relay conducts RTX of s_2. If U1 fails to decode s_1 after this RTX, then the best relay conducts RTX of s_1. If U1 and U2 fails to decode s_2, then the best relay from D_2 conducts RTX of s_2 to U1 (U2 can also receive this RTX). If U1 fails to decode s_1 after the RTX, then the best relay from D_1 conducts RTX of s_1. If the RTXs fail or D_1/D_2 are empty, then the eNodeB sends a new $s_{1,2}$ to U1, U2 and all R_i, rather than continuing to conduct RTX. U2 will also receive any RTX of s_2, which is intended for U1.

However, this scheme is impractical for MMC/IoT, because each device pair requires a large number of relays to increase the probability of successful decoding with no RTX or 1 RTX.

5.2.4 Cooperative Relaying NOMA

For the example 2-device NOMA cluster shown in Figure 5.5, UE1 can act as a relay for UE2 to increase UE2's reception ability, since UE1 decodes UE2's message during the SIC process [54]. This cooperative relay technique is termed 'decode-and-forward'

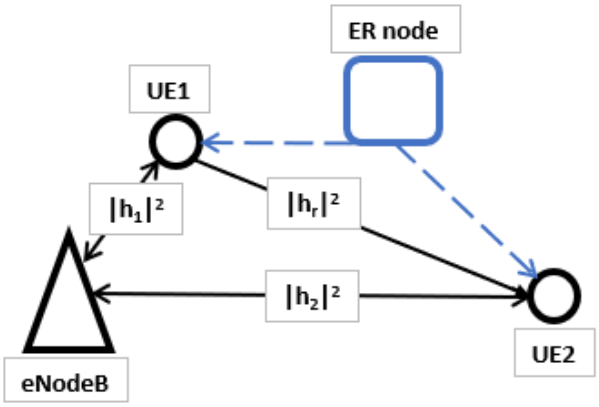


Figure 5.5: A 2-device NOMA cluster, with an ER node to supply power for UL transmission. **NOTE: The solid lines signify data transmission (and PS SWIPT in the DL) and the dashed lines signify EH harvesting for the UL.**

(DF) [57]. In cooperative NOMA, UE1 sends the UE2 post-SIC signal towards UE2, with the resulting SINR of UE2 given in Equation (5.9), where P_r is the transmit power of UE1 (when acting as a relay), and $|h_r|^2$ is the Rayleigh fading channel gain between UE2 and UE1.

$$\text{SINR}_{[2+2]\rightarrow 2} = \frac{P_2 |h_2|^2}{N_0 B + P_1 |h_2|^2} + \frac{P_r |h_r|^2}{N_0 B} \tag{5.9}$$

Maximal Ratio Combining (MRC) is used to combine SINR_2 with SNR_2 [129]. SINR_2 is the SINR at UE2 to decode $x_2(t)$ [assuming perfect SIC], as given earlier in Equation (3.5).

5.3 System Model

An example system model of $n = \{1, 2, 3, \ldots, d\}$ M2M devices designated into $m = \{1, 2, 3, \ldots, k\}$ clusters, by using the enhanced K-means Clustering scheme in Chapter 4, for the formation of clusters, was shown previously in Figure 3.1. Clustering is used to significantly reduce the complexity of SIC (and therefore power consumption) and the complexity of coordinating user cooperation in cooperative NOMA [53].

A simple 2-device NOMA cluster (UE1/UE2 pair) as shown in Figure 5.5, is considered to simplify the analysis of the proposed NOMA HARQ scheme, since the same scheme will be used for all clusters. As proposed in Section 4.7, an ER node will be used to simplify the EH process. The ER node will broadcast an EH signal (this will occur twice

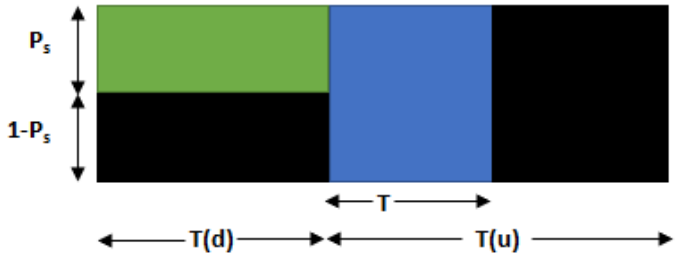


Figure 5.6: The normalised transmission time is split into 2 parts of $T(d)$ and $T(u)$, for DL NOMA with PS SWIPT and UL NOMA with TS SWIPT, respectively. P_s is the portion of the received normalised signal used for EH.

if HARQ and MRC soft combining is required) in between the DL and UL portions of the normalised transmission time, so that UL transmission is feasible for all devices. Each device will use PS SWIPT to power the DL SIC NOMA process (i.e. devices will use a portion of the received RF signal for EH) as in Chapter 3.

5.4 Enhanced NOMA HARQ

It was proposed in Section 4.7 that the normalised transmission time is split into two parts (as shown in Figure 5.6), where $1 = T(d) + T(u)$. $T(d)$ consists of DL NOMA (as in **Algorithm 3.2**) and accompanied by PS SWIPT, and $T(u)$ consists of UL NOMA (as in **Algorithm 4.1**) and accompanied by TS SWIPT. $T(u)$ is further split into 2 sub-parts, with the first sub-part (i.e. T) consisting of EH from the ER node, and the second sub-part consisting of UL transmission.

There are 4 SIC decoding scenarios for the example 2-device NOMA cluster shown in Figure 5.5, where the signals of interest are $x_1(t)$ and $x_2(t)$ for UE1 and UE2, respectively. These scenarios are given below (shown visually in **Table 5.1**):

1. **No RTX is required**: UE1 and UE2 correctly decode $x_1(t)$ and $x_2(t)$, respectively.
2. **RTX is required for UE1**: UE2 correctly decodes $x_2(t)$, but UE1 incorrectly decodes:
   (a) $x_2(t)$.
   (b) or $x_1(t)$ [implying $x_2(t)$ was decoded correctly].

Table 5.1: 2-device SIC decoding scenarios, where • indicates SIC decoding success and × indicates failure.

Scenario	UE1		UE2	HARQ?	Wrong signals?
	$x_1(t)$	$x_2(t)$			
1	•	•	•	No	0
2a		×	•	UE1	1
2b	×	•	•	UE1	1
3	•	•	×	UE2	1
4a		×	×	UE1, UE2	2
4b	×	•	×	UE1, UE2	2

3. **RTX is required for UE2**: UE1 correctly decodes $x_1(t)$, but UE2 incorrectly decodes $x_2(t)$.

4. **RTX is required for UE1 and UE2**: UE2 incorrectly decodes $x_2(t)$, and UE1 incorrectly decodes:

   (a) $x_2(t)$.

   (b) or $x_1(t)$ [implying $x_2(t)$ was decoded correctly].

Based on **Table 5.1**, the '1 wrong' signal is the more dominant scenario, which agrees with the findings in [122]. The following descriptions of the HARQ process (assuming ACK/NAK is error-free and the latency is negligible [129]) for scenario 1–4b in **Table 5.1**, are given below, where the eNodeB performs RTX with one of the following 3 RVs: (**i**) 1 RV consisting of the superimposed signal, labelled as RV(x); (**ii**) 1 RV consisting of $x_1(t)$, labelled as RV(x1) and; (**iii**) 1 RV consisting of $x_2(t)$, labelled as RV(x2); for which the device that sent a NAK applies packet combining with the erroneous packet stored in the soft buffer:

1. For scenario 1, UE1 and UE2 send an ACK to the eNodeB, while UE2 also sends an ACK to UE1.

2. For scenario 2a and 2b, UE2 sends an ACK to UE1 and the eNodeB, while UE1 sends a NAK to the eNodeB. The eNodeB sends RV(x) to UE1 for scenario 2a, and RV(x1) for scenario 2b.

3. For scenario 3, UE1 sends an ACK to the eNodeB, while UE2 sends a NAK to UE1 and the eNodeB. UE1 then performs cooperative relaying by sending the UE2 post-SIC signal to UE2. The eNodeB also sends RV(x2) to UE2.

4. For scenario 4a, UE1 and UE2 send a NAK to the eNodeB, while UE2 also sends a NAK to UE1. The eNodeB sends RV(x) to both UE1 and UE2.

5. For scenario 4b, UE2 sends a NAK to UE1 and the eNodeB, while UE1 sends a NAK to the eNodeB after UE1 sends the UE2 post-SIC signal to UE2, by performing cooperative relaying. The eNodeB sends RV(x) to both UE1 and UE2.

The steps for the proposed enhanced NOMA HARQ scheme is given in **Algorithm 5.2**. Note that **Algorithm 5.2** is only applicable to 2-device NOMA clusters, but can be generalised to y-sized clusters, with $y(y-1)+2$ decoding scenarios (i.e. 8 for $y=3$ and 14 for $y=4$).

Algorithm 5.2: Enhanced HARQ in NOMA

1. The network of M2M devices are arranged into k clusters by using the enhanced K-means clustering algorithm from **Algorithm 4.1**.

2. DL NOMA and PS SWIPT are performed.

3. If a NAK is received from UE1 and/or UE2, the eNodeB performs RTX of one of the 3 RVs as described on page 106 for scenarios 2a-4b. For scenarios 3 and 4b, UE1 performs cooperative relaying NOMA after it has decoded UE2's signal during the SIC process.

4. While UE1 and UE2 are attempting to decode the superimposed signal in step 2, the ER node broadcasts an EH signal to UE1 and UE2. If a NAK is received from UE1 and/or UE2, the ER node broadcasts another EH signal to UE1 and UE2.

5. Once UE1 and UE2 have decoded the intended signals (either initially or by packet combining in the soft buffer), then UL NOMA is performed, where the energy harvested in step 4 is utilised.

5.5 Outage Probability, SINR and Achievable Rates

5.5.1 OMA/NOMA Outage Probabilities

As given in Subsubsection 5.2.3.2, the outage (defined as "when the achievable rate is lower than the target rate" [131]) probability[3] of OMA (while using CC-HARQ) is shown in Equation (5.10), where T is the number of transmission rounds (i.e. initial transmission + further RTX), $\hat{R}_j$ is the target rate of the j-th device, and SNR_j^t is the SNR of the j-th device during the t-th transmission round [79].

$$P_{\text{out}}\left(^{\text{OMA}}_{j,T,\text{CC}}\right) = \mathrm{P}\left\{\frac{1}{2}\log_2\left(1+\sum_{t=1}^{T}\mathrm{SNR}_j^t\right) \leq T\hat{R}_j\right\} \tag{5.10}$$

The outage probability of OMA (while using IR-HARQ) is shown in Equation (5.11) [123].

$$P_{\text{out}}\left(^{\text{OMA}}_{j,T,\text{IR}}\right) = \mathrm{P}\left\{\frac{1}{2}\sum_{t=1}^{T}\log_2\left(1+\mathrm{SNR}_j^t\right) \leq T\hat{R}_j\right\} \tag{5.11}$$

The outage probabilities for UE1 and UE2 (while using CC-HARQ) are shown in Equations (5.12) and (5.13) [79]. E.g. $\mathrm{SINR}_{1\to 2}^t$ is the SINR at UE1 to decode UE2's signal.

$$\begin{aligned} P_{\text{out}}\left(^{\text{NOMA}}_{1,T,\text{CC}}\right) = 1 - \mathrm{P}\Bigg\{&\log_2\left(1+\sum_{t=1}^{T}\mathrm{SINR}_{1\to 2}^t\right) > T\hat{R}_2, \\ &\log_2\left(1+\sum_{t=1}^{T}\mathrm{SINR}_{1\to 1}^t\right) > T\hat{R}_1\Bigg\} \end{aligned} \tag{5.12}$$

$$P_{\text{out}}\left(^{\text{NOMA}}_{2,T,\text{CC}}\right) = \mathrm{P}\left\{\log_2\left(1+\sum_{t=1}^{T}\mathrm{SINR}_{2\to 2}^t\right) \leq T\hat{R}_2\right\} \tag{5.13}$$

The outage probability for UE1 and UE2 (while using IR-HARQ) are shown in Equations (5.14) and (5.15) [123].

$$\begin{aligned} P_{\text{out}}\left(^{\text{NOMA}}_{1,T,\text{IR}}\right) = 1 - \mathrm{P}\Bigg\{&\sum_{t=1}^{T}\log_2\left(1+\mathrm{SINR}_{1\to 2}^t\right) > T\hat{R}_2, \\ &\sum_{t=1}^{T}\log_2\left(1+\mathrm{SINR}_{1\to 1}^t\right) > T\hat{R}_1\Bigg\} \end{aligned} \tag{5.14}$$

$$P_{\text{out}}\left(^{\text{NOMA}}_{2,T,\text{IR}}\right) = \mathrm{P}\left\{\sum_{t=1}^{T}\log_2\left(1+\mathrm{SINR}_{2\to 2}^t\right) \leq T\hat{R}_2\right\} \tag{5.15}$$

[3]Note that the analytical results of the outage probabilities shown in Subsection 5.5.1, were already evaluated in [79] and [123].

5.5.2 SINRs and Achievable Rates of the Table 5.1 scenarios

For scenario 2a, the SINR at UE1 after receiving the RTX of RV(x) from the eNodeB is given in Equation (5.16), assuming UE1 correctly decodes $x_1(t)$ after RTX.

$$\mathrm{SINR}_{[1+\mathrm{RV(x)}]\to 1} = \frac{P_2 \left|h_1\right|^2}{N_0 B + P_1 \left|h_1\right|^2} + \frac{P_1 \left|h_1\right|^2}{N_0 B} \tag{5.16}$$

For scenario 2b, the SINR at UE1 after receiving the RTX of RV(x1) from the eNodeB at full power (i.e. $P_1 + P_2$) is given in Equation (5.17). It is implied UE1 decoded $x_2(t)$ correctly but failed to decode $x_1(t)$, leading to RTX.

$$\mathrm{SINR}_{[1+\mathrm{RV(x1)}]\to 1} = \frac{P_1 \left|h_1\right|^2}{N_0 B} + \frac{P_t \left|h_1\right|^2}{N_0 B} \tag{5.17}$$

For scenario 3, the SINR at UE2 after receiving the RTX of RV(x2) from the eNodeB at full power and RTX of the post-SIC UE2 signal from UE1, is given in Equation (5.18).

$$\mathrm{SINR}_{[2+\mathrm{RV(x2)+rel}]\to 2} = \frac{P_2 \left|h_2\right|^2}{N_0 B + P_1 \left|h_2\right|^2} + \frac{P_r \left|h_r\right|^2}{N_0 B} + \frac{P_t \left|h_2\right|^2}{N_0 B} \tag{5.18}$$

For scenario 4a, the SINR at UE1 after receiving the RTX of RV(x) from the eNodeB is given in Equation (5.16), while the SINR at UE2 after receiving the RTX of RV(x) from the eNodeB is given in Equation (5.19).

$$\mathrm{SINR}_{[2+\mathrm{RV(x)}]\to 2} = 2\left[\frac{P_2 \left|h_2\right|^2}{N_0 B + P_1 \left|h_2\right|^2}\right] \tag{5.19}$$

For scenario 4b, the SINR at UE1 after receiving the RTX of RV(x) from the eNodeB is given in Equation (5.20), while the SINR at UE2 after receiving the RTX of RV(x) from the eNodeB is given in Equation (5.21).

$$\mathrm{SINR}_{[1+\mathrm{RV(x)}]\to 1} = 2\left[\frac{P_1 \left|h_1\right|^2}{N_0 B}\right] \tag{5.20}$$

$$\mathrm{SINR}_{[2+\mathrm{RV(x)+rel}]\to 2} = 2\left[\frac{P_2 \left|h_2\right|^2}{N_0 B + P_1 \left|h_2\right|^2}\right] + \frac{P_r \left|h_r\right|^2}{N_0 B} \tag{5.21}$$

The achievable rates for each of the scenarios are simply calculated as $\log_2(1+\sum_{t=1}^{T} \mathrm{SINR}^{\mathrm{t}})$ for CC-HARQ, and $\sum_{t=1}^{T} \log_2(1 + \mathrm{SINR}^{\mathrm{t}})$ for IR-HARQ.

Table 5.2: MATLAB HARQ (2-device NOMA cluster) Simulation Parameters

Parameters	Setup
Bandwidth	360 kHz
Number of available resource blocks	2
Transmit Antennas	1
Receive Antennas	1
DL Transmit Power/cluster, P_t	1.5924 W [117]
UE1 DL Transmit Power, P_1	$0.3P_T$ W
UE2 DL Transmit Power, P_2	$0.7P_T$ W
Path loss exponent, α	3.76 [117]
UE1 Minimum rate requirement, $\hat{R}_1$	288 kbps [79]
UE2 Minimum rate requirement, $\hat{R}_2$	72 kbps [79]
Transmission rounds, T	1 and 2 [79]

5.6 Results and Discussion

As stated in Section 5.3, a simple 2-device NOMA cluster (UE1/UE2 pair) is considered to simplify the analysis of the proposed NOMA HARQ scheme. The simulation parameters shown in **Table 5.2** were used for NOMA and OMA. For each cluster, $P_t = (\omega P_T \ / \ 50)$, where $P_T = 46$ dBm $= 39.8107$ W, for an eNodeB in LTE-A, and for each device, $P_t' = 24$ dBm $= 0.25$ W [117].

5.6.1 Outage Probability

5.6.1.1 Comparison between $T = 1$ and $T = 2$

A comparison of the OMA/NOMA outage probabilities for UE1 and UE2 with $T = 1$ (no RTX) and $T = 2$ (using CC-HARQ or IR-HARQ) is shown in Figure 5.7. Note that for UE1, the outage probability is based on decoding its own signal $x_1(t)$, after correctly decoding $x_2(t)$ during the SIC process, for both $T = 1$ and $T = 2$.

It can be observed in Figure 5.7 that when $T = 1$, the outage probability of NOMA UE2 is better than OMA UE2 when the transmit SNR, ρ, is $11 \leq \rho \leq 40$ dB ($\approx 95\%$ better at 40 dB); while the outage probability of NOMA UE1 is worse than OMA UE1 for when $12 \leq \rho \leq 40$ dB ($\approx 21\%$ worse at 40 dB). When $T = 2$, the IR-HARQ outage

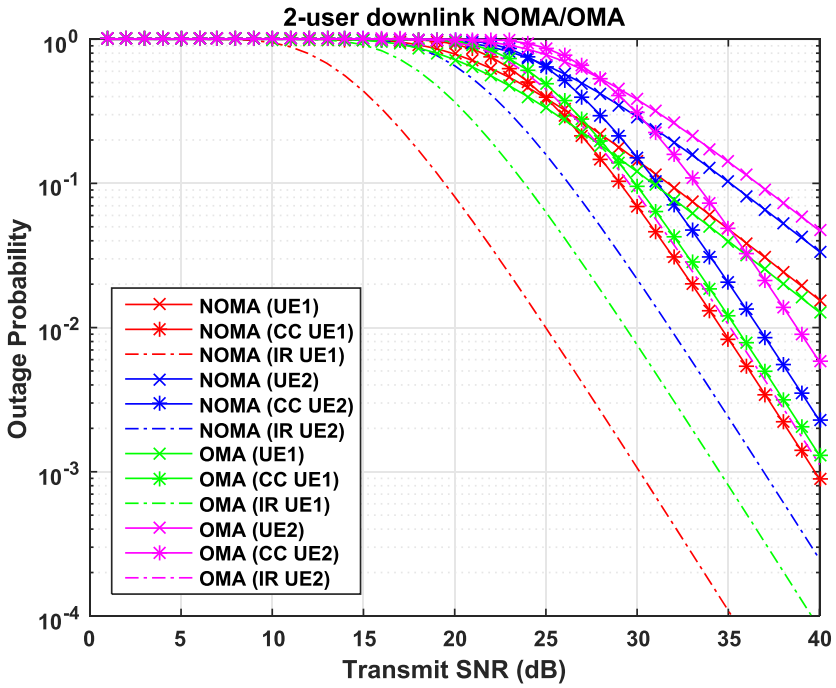


Figure 5.7: 2-device outage probability UE1/UE2 comparison between NOMA and OMA, for $T = 1$ and $T = 2$.

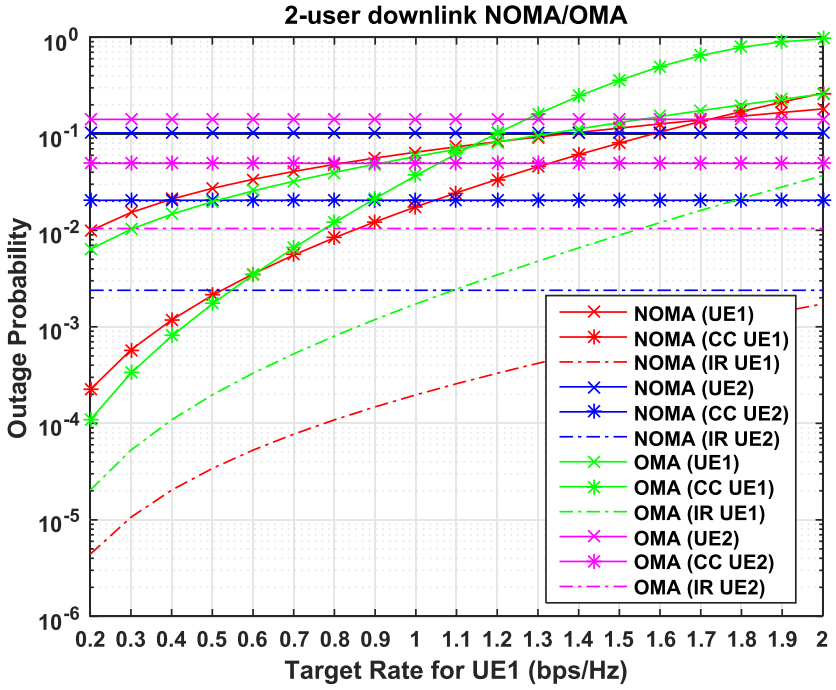


Figure 5.8: 2-device outage probability UE1/UE2 comparison between NOMA and OMA, with varying $\hat{R}_1$.

probabilities are clearly superior to the CC-HARQ outage probabilities, while the CC-HARQ outage probabilities are better than the $T = 1$ outage probabilities. However, the CC-HARQ outage probability is worse for NOMA UE1 when $12 \leq \rho \leq 25$ dB, for NOMA UE2 when $16 \leq \rho \leq 24$ dB, for OMA UE1 when $12 \leq \rho \leq 28$ dB, and for OMA UE2 when $17 \leq \rho \leq 27$ dB.

In Figure 5.8, with $\rho = 35$ dB, $\hat{R}_2 = 72$ kbps (0.2 bps/Hz), and $72 \leq \hat{R}_1 \leq 720$ kbps ($0.2 \leq \hat{R}_1 \leq 2$ bps/Hz), the outage probability for UE1 is worse than UE2, when $1.3 < \hat{R}_1 \leq 2$ bps/Hz for NOMA $T = 0$, when $1 < \hat{R}_1 \leq 2$ bps/Hz for NOMA CC-HARQ, when $1.5 < \hat{R}_1 \leq 2$ bps/Hz for OMA $T = 0$, when $1 < \hat{R}_1 \leq 2$ bps/Hz for OMA CC-HARQ, and $1.5 < \hat{R}_1 \leq 2$ bps/Hz for OMA IR-HARQ.

5.6.1.2 Comparison between Table 5.1 SIC decoding scenarios

A comparison of the outage probabilities for each scenario (labelled as 'Sc') from **Table 5.1**, is shown in Figure 5.9. It can be observed that the CC-HARQ outage probability for UE1, under scenarios 2b and 4b, are exactly the same as CC-HARQ in Figure 5.7, since UE1 has successfully decoded its own signal $x_1(t)$, after correctly decoding $x_2(t)$

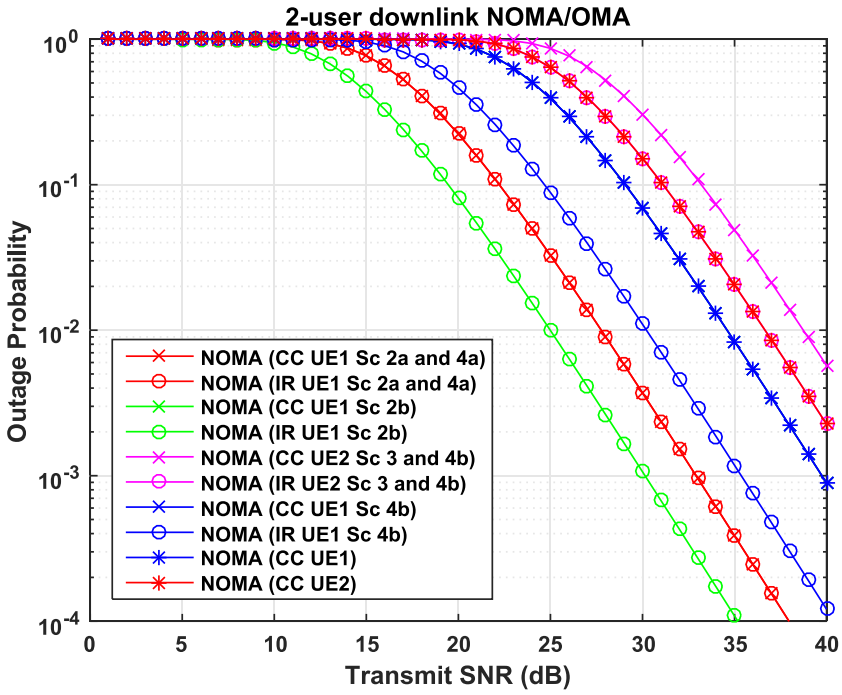


Figure 5.9: 2-device outage probability UE1/UE2 comparison with proposed NOMA-HARQ scheme.

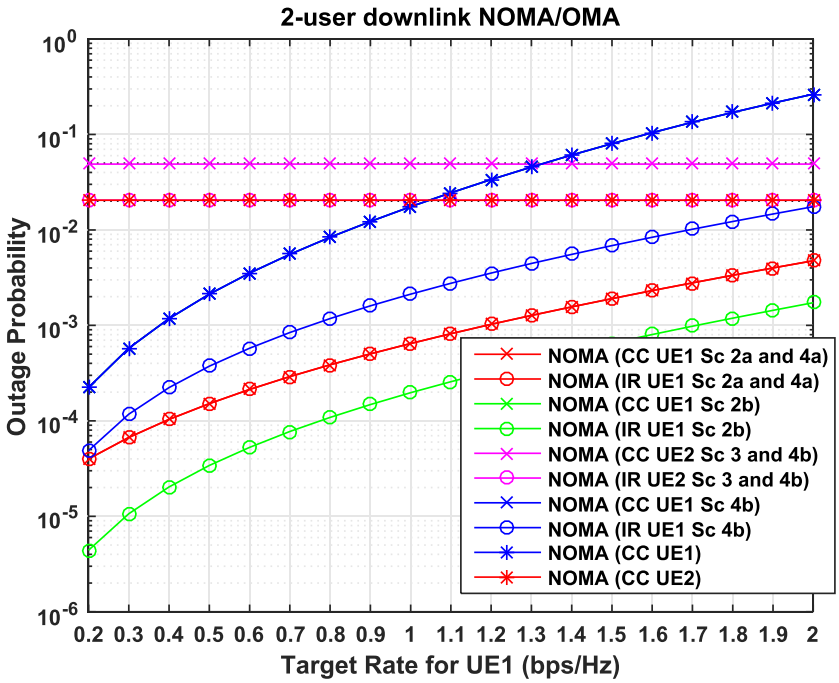


Figure 5.10: 2-device outage probability UE1/UE2 comparison with proposed NOMA-HARQ scheme, with varying $\hat{R}_1$.

from the resulting signal of the MRC process. CC-HARQ and IR-HARQ outage probability for UE1, under scenarios 2a and 4a, are also exactly the same, as UE1 receives the same RV(x) from the eNodeB. The IR-HARQ outage probability for UE1, under scenario 4b, is superior to CC-HARQ, and also superior for UE2, under scenarios 3 and 4b.

In Figure 5.10, with $\rho = 35$ dB, $\hat{R}_2 = 72$ kbps (0.2 bps/Hz), and $72 \leq \hat{R}_1 \leq 720$ kbps ($0.2 \leq \hat{R}_1 \leq 2$ bps/Hz), the outage probability for UE1 is worse than UE2, when $1.3 < \hat{R}_1 \leq 2$ bps/Hz for scenario 4b. Otherwise, UE1 scenarios 2a, 2b, 4a and 4b follow the same trend as shown in Figure 5.9. The UE1 CC-HARQ outage probability for scenarios 2b and 4b are worse than the UE2 CC-HARQ for scenarios 3 and 4b, when $1.3 < \hat{R}_1 \leq 2$ bps/Hz, and worse than UE2 IR-HARQ for scenarios 3 and 4b, when $1 < \hat{R}_1 \leq 2$ bps/Hz.

5.6.2 Achievable Rates

A comparison of the achievable rates for each scenario (labelled as 'Sc') from **Table 5.1**, is shown in Figure 5.11. Each scenario is also compared to the CC-HARQ NOMA/OMA cases from [79] and to OMA for $T = 1$. For UE1, scenario 2b achieves the best achievable

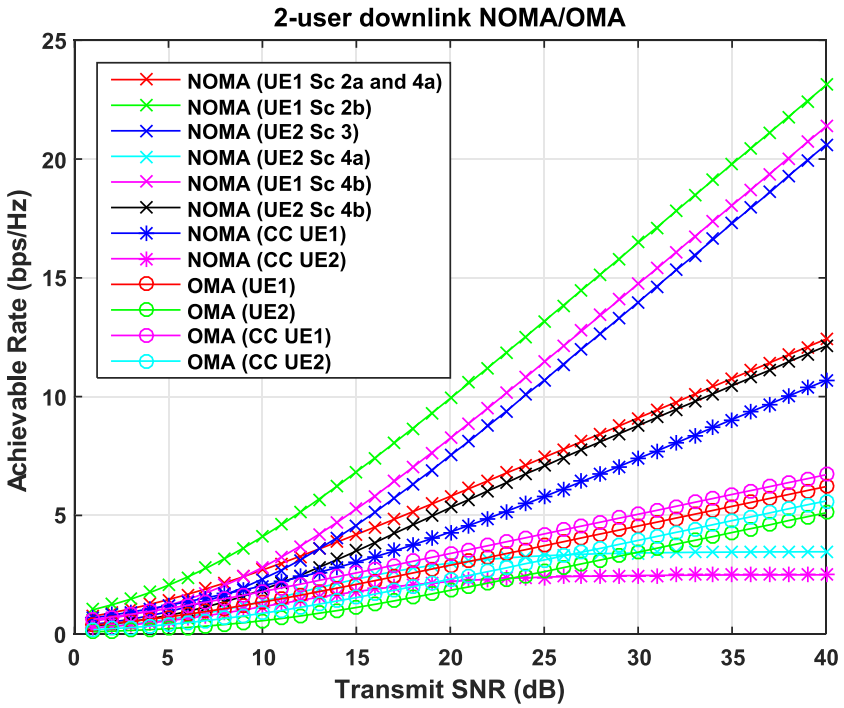


Figure 5.11: 2-device achievable rate UE1/UE2 comparison with proposed NOMA-HARQ scheme. The proposed scheme is also compared against OMA for $T = 1$ and CC-HARQ.

rate, because UE1 failed to decode $x_1(t)$, and UE1 received RV(x1) at full power from the eNodeB [since UE2 successfully decoded $x_2(t)$]. For UE2, scenario 3 achieves the best achievable rate, because UE2 received RV(x2) at full power from the eNodeB and also the post-SIC UE2 signal from UE1 during cooperative relaying NOMA. Most of the scenarios using the proposed scheme, have better achievable rates than the CC-HARQ cases from [79], except for UE2 during scenario 4a.

5.7 Conclusion

In this chapter, an enhanced NOMA scheme with HARQ was proposed to fulfil the requirements of URLLC devices (such as mission critical industrial control, medical, and V2X applications), which have the QoS requirements of ultra-reliability and ultra-low latency. HARQ was used because it has been proven to improve the reliability of data transmission in cellular networks. Cooperative relaying NOMA was also used to increase the reception ability and provide RTX diversity at the weak channel device. An ER node

was used to perform TS SWIPT instead of the eNodeB. IR-HARQ was used as the basis for the proposed scheme, because it has been shown to achieve a higher spectral efficiency than CC [127].

The proposed scheme used the same enhanced K-means clustering algorithm as in **Algorithm 4.1** of Chapter 4, since there were some throughput problems observed in the UL (if the proposed scheme is extended to clusters of larger than size 3)[4]. DL NOMA and PS SWIPT were then performed to power the SIC process at UE1. While UE1 and UE2 are attempting to decode their signals of interest, the ER node sends an EH signal to UE1 and UE2 to power the future UL transmission. Note that simultaneously transmitting and EH at a device is infeasible [56], not when simultaneously receiving and EH. The eNodeB performs RTX with one of the 3 RVs for scenarios 2a-4b. UE1 performs cooperative relaying NOMA by sending the UE2 post-SIC signal to UE2 for scenarios 3 and 4b. UL NOMA is then performed after UE1 and UE2 have decoded their intended signals, using the energy previously harvested.

The proposed scheme was shown in Figure 5.9, to have the same UE1 outage probability as NOMA IR-HARQ for scenario 2b, which is not surprising since it used IR-HARQ as the basis. The proposed scheme had a significant outage probability gain over the CC-HARQ scheme used in [79] and the LTE-A OMA scheme. It was also shown in Figure 5.11 that the proposed scheme achieved a significant achievable rate gain over CC-HARQ and the LTE-A OMA scheme. Scenario 2b achieved the highest achievable rate for UE1, while it was scenario 3 for UE2. This is due to the eNodeB sending RV(x1) and RV(x2) to UE1 and UE2, respectively, at full transmission power, rather than RV(x) which is sent with the usual NOMA split power as the initial transmission.

The proposed scheme facilitates the URLLC operating mode while taking into account that each device's battery is finite. Incorporating cooperative relaying NOMA also improved the SINR at UE2 for scenarios 3 and 4b, because UE2 receives the post-SIC UE2 signal from UE1. However, the SINR for scenario 3 is better than in scenario 4b, because UE2 receives RV(x2) at full transmission power from the eNodeB. Applying TS SWIPT by using an ER node instead of the eNodeB, ensured that the power consumption of the future UL transmission, of performing packet combining, and cooperative relaying NOMA, was feasible, without having to sacrifice more time slots at the eNodeB.

[4]The authors in [76] did not consider cases with normalised channel gain differences of < 1.5 dB (see Table 5 of [76]), meaning that they never observed this UL throughput problem.

Chapter 6

Thesis Conclusions and Future Work

6.1 Conclusions

The main focus of this thesis was to enable the 5G MMC operating mode, by addressing the QoS requirements of massive connectivity, small packets, low data rate, high EE and low energy consumption. Sensor nodes were used as motivations in Chapters 3 and 4, since the IoT will mostly consist of these devices. 'User Pairing' ensured that each cluster had a strong channel device and also prevented the final clustering solution featuring 1-device clusters. This enhanced the NOMA network sum throughput over the current LTE-A OMA techniques, compared to the traditional K-means and Hierarchical clustering algorithms. SWIPT EH techniques were incorporated to address the DL SIC and UL transmission power consumption, which led to increased EE.

This thesis also had a secondary focus of enabling the 5G URC operating mode, by addressing the stricter (compared to LTE-A) QoS requirements of ultra-reliability and ultra-low latency (i.e. URLLC). Mission critical industrial control, medical, and V2X applications were used as motivations in Chapter 5, to improve the RTX process and therefore reduce the delay from the possible multiple RTX.

6.1.1 DL NOMA with SWIPT

An enhanced K-means clustering algorithm accompanied with DL NOMA and PS SWIPT was proposed in Chapter 3, to fulfil the QoS requirements for MMC devices as mentioned in Section 6.1. MMC devices will likely suffer from the *overload and access problem*, if the

OMA technique is used. Clustering was used to reduce the feedback and control signalling overhead, while NOMA was used to avoid the *overload and access problem*, which fulfils objective **O1** (i.e. propose a clustering algorithm with NOMA to address the possible simultaneous massive random access, and to avoid the *overload and access problem*).

Clustering was also used to reduce the complexity (and therefore power consumption) of the NOMA SIC process for an un-clustered MMC network. In order to address the EE and energy consumption of SIC, the PS SWIPT EH technique was used to mitigate the power consumption of SIC, to fulfil objective **O3** (i.e. mitigate the power consumption of DL SIC by using SWIPT). Therefore, this established the essential link between clustering and NOMA for MMC/IoT.

The proposed scheme involved determining the optimal value of k clusters by using the 'silhouette value' metric and 'Sum of Squared Error' metric, and then the K-means clustering algorithm was applied to the network. The k strongest channel gain devices were excluded from the network and were then reassigned to the appropriate cluster to ensure that each cluster had a strong channel gain device, in accordance with 'User Pairing', which fulfils objective **O2** (i.e. incorporate user pairing to enhance the NOMA network sum throughput gain over OMA). Power was optimally allocated to each device within a cluster based on a set of equations and corresponding necessary conditions.

The proposed scheme was shown to have a higher NOMA cluster sum throughput over OMA for each cluster formed. It was also shown to have a 34.94% higher average NOMA network sum throughput (over 25 simulation runs) over OMA, a 13.47% improvement over traditional K-means clustering and a 22.00% improvement over Hierarchical clustering, at a minimum rate requirement of $R_i = 100$ kHz.

For 2-device clusters, it was shown that DL SIC is feasible (i.e. the portion of the received RF signal that was sacrificed for EH did not lead to outage), with a slight decrease in the NOMA sum throughput. NOMA also had a superior sum throughput over OMA, when the portion of the received normalised RF signal used for EH, P_s, was ≥ 0.46. Extending to 3-device clusters, it was shown that DL SIC is also feasible, which is a promising result for utilising DL SIC for cluster sizes of larger than 2. Applying PS SWIPT also showed promising results in increasing the EE by mitigating the power consumption of SIC.

6.1.2 UL NOMA with SWIPT

An enhanced K-means clustering algorithm accompanied with UL NOMA and TS SWIPT was proposed in Chapter 4, to fulfil the QoS requirements of MMC devices as given in Section 6.1. However, an additional step was required for the UL K-means clustering algorithm (which already fulfilled objectives **O1** and **O3**, as detailed in Subsection 6.1.1), as an outage would likely occur if the difference between the channel gains of the second strongest device and the 2 weakest devices in a 4-device cluster was too small.

In order to address the EE and energy consumption of UL device transmission, the TS SWIPT EH technique was used to mitigate the power consumption of UL device transmission, to fulfil objective **O3** (i.e. mitigate the power consumption of UL device transmission by using SWIPT). Note that the power consumption of SIC is not as significant in the UL, as the eNodeB has much more power available to perform SIC. Nevertheless, the power consumption of UL SIC was automatically reduced with the additional step of reassigning the second strongest device in every 4-device cluster to the nearest 2-device cluster.

The proposed scheme was shown to have a higher NOMA cluster sum throughput over OMA for each cluster formed. It was also shown to have a 24.77% higher average NOMA network sum throughput (over 25 simulation runs) over OMA, a 24.06% improvement over traditional K-means clustering and a 24.49% improvement over Hierarchical clustering, at $R_i = 100$ kHz.

For 2-device clusters, it was shown that UL device transmission is feasible (assuming that each device's battery is finite), with a slight decrease in the NOMA sum throughput. This was when the portion of the normalised transmission time period, $T_s = 0.1$. It was also shown that UL device transmission is feasible for $0.2 < T_s < 0.92$, with $|h_1|^2 = 40$ dB and $|h_2|^2 = 30$ dB, at $R_i = 100$ kHz. Extending to 3-device clusters, it was shown that UL device transmission is also feasible, which is a promising result for utilising UL transmission for cluster sizes of larger than 2. Applying TS SWIPT (without an MTCG) also showed promising results in increasing the EE by mitigating the power consumption of UL device transmission.

6.1.3 Enhanced NOMA HARQ Scheme

An enhanced NOMA HARQ scheme was proposed in Chapter 5 accompanied with hybrid PS/TS SWIPT, to fulfil the QoS requirements for URLLC devices as mentioned in

Section 6.1. HARQ was used because it was proven to improve the reliability of data transmission in cellular networks. Cooperative relaying NOMA was also used to increase the reception ability of the weak channel device and for RTX diversity if a NAK is received from the weak channel device. An ER node was also used to perform TS SWIPT instead of the eNodeB. IR-HARQ was used as the basis for the enhanced NOMA scheme with HARQ, because it has been shown to achieve a higher spectral efficiency than CC.

The proposed scheme used the same enhanced K-means clustering algorithm as in Chapter 4. For an example 2-device cluster (UE1/UE2 device pair, where UE1 is the strong channel gain device) DL NOMA and PS SWIPT was then performed to mitigate the power consumption of the SIC process at UE1. While UE1 and UE2 are attempting to decode their signals of interest of $x_1(t)$ and $x_2(t)$, the ER node sends an EH signal to UE1 and UE2 to power the future UL transmission. Note that simultaneously receiving and EH at a device is feasible [56]. The eNodeB performs RTX with one of the 3 RVs [i.e. a RV of the superimposed signal, RV(x); a RV of UE1's signal, RV(x1) and a RV of UE2's signal, RV(x2)] for Table 5.1 scenarios 2a-4b. UE1 performs cooperative relaying NOMA by sending the UE2 post-SIC signal to UE2 for scenarios 3 and 4b. UL NOMA is then performed after UE1 and UE2 have successfully decoded $x_1(t)$ and $x_2(t)$, respectively, by using the energy previously harvested.

The proposed scheme was shown to have the same UE1 outage probability as NOMA IR-HARQ for scenario 2b. The proposed scheme had a significant outage probability gain over the competing CC-HARQ scheme. It was also shown that the proposed scheme achieved a significant achievable rate gain over CC-HARQ. Scenario 2b achieved the highest achievable rate for UE1, while scenario 3 achieved the highest achievable rate for UE2. This is due to the eNodeB sending RV(x1) and RV(x2) to UE1 and UE2, respectively, at full transmission power (i.e. P_1+P_2), rather than RV(x) which is sent with the initial transmission power. Incorporating cooperative relaying NOMA in scenarios 3 and 4b, also improved the achievable rate because UE2 received the post-SIC UE2 signal from UE1.

The proposed scheme facilitates the URLLC operating mode, by fulfilling objective **O4** (i.e. using HARQ to ensure the reliability of data transmission and therefore also reduce delay) while taking into account that each device's battery is finite. Applying TS SWIPT by using an ER node instead of the eNodeB, ensured that the power consumption of: (**i**) the future UL transmission, (**ii**) performing packet combining, and; (**iii**) cooperative

relaying NOMA; was feasible, without having to sacrifice more time slots at the eNodeB.

6.2 Future Work

6.2.1 Non-Orthogonal Waveforms

Many of the related literature featured in this thesis, and the proposed solutions in Chapters 3, 4 and 5, utilise the current OFDM waveform. Non-Orthogonal waveforms have been proposed for 5G cellular networks such as Filter-bank based Multicarrier (FBMC), Universal Filtered Multi-Carrier (UFMC), and Generalised Frequency Division Multiplexing (GDFM) [132–136].

The advantages of these waveforms are: they have reduced sidelobes compared with OFDM, CP and guard intervals (which are used for TDD and FDD) are no longer required, interference between adjacent subcarriers is flexible, and spectrum resources can be scattered. Generally, these proposed non-orthogonal waveforms could replace OFDM waveforms if the inter-carrier interference and the Rx complexity are reduced.

6.2.2 MIMO

MIMO has been shown to provide improvements in spectral efficiency and also in error probability, especially when it is combined with NOMA [51,137]. MIMO-NOMA has been shown to be superior to MIMO-OMA, but there is an open problem of the heavy interuser interference in MIMO-NOMA that is not present in MIMO-OMA [51].

However, it was pointed out in [137], that the users' channel conditions are likely to be similar in practice, meaning that the performance gain of NOMA over OMA can be marginal. [137] has proposed to degrade a user's effective channel gain while also improving the signal strength at the other user. This solution to the problem of similar channel gains, would be useful to investigate in terms of the proposed UL NOMA scheme in Chapter 4, where the second strongest channel devices in each 4-device cluster were reallocated to neighbouring 2-device clusters instead.

MIMO can also be used for SWIPT [138] to increase the feasibility of the DL SIC process and UL transmission. Inter-user interferences in NOMA can instead be used for EH harvesting, so that it does not significantly affect information decoding [138].

6.2.3 Full-Duplex Communication

As mentioned in Subsection 2.5.5, full-duplex mode (i.e. 2 antennas for simultaneous UL and DL) can be used to improve the spectral efficiency of cooperative NOMA [107]. There are many other related fields that can benefit from full-duplex communication, such as HARQ [139], to further increase the reliability of data transmission in NOMA; and SWIPT [140], to further mitigate the power consumption (and further increase the EE) of the DL SIC process and UL transmission..

However, full-duplex mode suffers from 'self-interference', where the transmitted signal interferes with the received signal on the same transceiver [108–112]. Self-interference cancellation currently does not completely remove the self-interference, which results in self-interference residual at the device [108, 110, 111].

Bibliography

[1] E. Cabrera and R. Vesilo, "Enhanced K-means Clustering and Uplink Non-Orthogonal Multiple Access (NOMA) for Massive M2M Communication," *submitted to IEEE Internet of Things (IoT) Journal*, 2019.

[2] E. Cabrera and R. Vesilo, "An Enhanced K-means Clustering Algorithm with Non-Orthogonal Multiple Access (NOMA) for MMC Networks," in *2018 28th International Telecommunication Networks and Applications Conference (ITNAC)*, Nov 2018, pp. 1–8.

[3] E. Cabrera, G. Fang, and R. Vesilo, "Adaptive Hybrid ARQ (A-HARQ) for Ultra-Reliable Communication in 5G," in *2017 IEEE 85th Vehicular Technology Conference (VTC Spring)*, Jun 2017, pp. 1–6.

[4] E. Cabrera, "Adaptive HARQ (A-HARQ) for Ultra-Reliable Communication in 5G," *Master of Research Thesis*, Macquarie University, Sydney, Australia, 2015. [Online]. Available: http://hdl.handle.net/1959.14/1068214

[5] X. Tong, G. Fang, D. Nguyen, J. Lin, and E. Cabrera, "An Energy-Balanced Routing Algorithm in Wireless Seismic Sensor Network," *Journal of Computational and Theoretical Nanoscience*, vol. 13, no. 10, pp. 6823–6833, Oct 2016.

[6] S. Sesia, I. Toufik, and M. Baker, *LTE-The UMTS Long Term Evolution: From Theory to Practice*, 2nd ed. John Wiley & Sons, Ltd, 2011.

[7] "Accelerating the mobile ecosystem expansion in the 5G Era with LTE Advanced Pro," Qualcomm Technologies, Inc., May 2018. [Online]. Available: https://www.qualcomm.com/media/documents/files/accelerating-the-mobile-ecosystem-expansion-in-the-5g-era-with-lte-advanced-pro.pdf

[8] "Cisco Visual Networking Index: Forecast and Trends, 2017–2022 White Paper," Cisco Systems, Inc., Feb 2019. [Online]. Available: https://www.cisco.com/c/en/us/solutions/collateral/service-provider/visual-networking-index-vni/white-paper-c11-741490.pdf

[9] "Cisco Visual Networking Index: Global Mobile Data Traffic Forecast Update, 2017–2022 White Paper," Cisco Systems, Inc., Feb 2019. [Online].

Available: https://www.cisco.com/c/en/us/solutions/collateral/service-provider/visual-networking-index-vni/white-paper-c11-738429.pdf

[10] F. Boccardi, R. W. Heath, A. Lozano, T. L. Marzetta, and P. Popovski, "Five disruptive technology directions for 5G," *IEEE Communications Magazine*, vol. 52, no. 2, pp. 74–80, Feb 2014.

[11] C. Bockelmann, N. Pratas, H. Nikopour, K. Au, T. Svensson *et al.*, "Massive machine-type communications in 5g: physical and MAC-layer solutions," *IEEE Communications Magazine*, vol. 54, no. 9, pp. 59–65, Sep 2016.

[12] H. D. Schotten, R. Sattiraju, D. G. Serrano, Z. Ren, and P. Fertl, "Availability indication as key enabler for ultra-reliable communication in 5G," in *2014 European Conference on Networks and Communications (EuCNC)*, Jun 2014, pp. 1–5.

[13] A. Ijaz, L. Zhang, M. Grau, A. Mohamed, S. Vural *et al.*, "Enabling Massive IoT in 5G and Beyond Systems: PHY Radio Frame Design Considerations," *IEEE Access*, vol. 4, pp. 3322–3339, Jun 2016.

[14] S. Li, L. D. Xu, and S. Zhao, "5G Internet of Things: A survey," *Journal of Industrial Information Integration*, vol. 10, pp. 1–9, 2018. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S2452414X18300037

[15] G. A. Akpakwu, B. J. Silva, G. P. Hancke, and A. M. Abu-Mahfouz, "A Survey on 5G Networks for the Internet of Things: Communication Technologies and Challenges," *IEEE Access*, vol. 6, pp. 3619–3647, Jan 2018.

[16] "5G is rolling out globally — and faster than any G before it," Qualcomm Technologies, Inc., Jun 2019. [Online]. Available: https://www.qualcomm.com/news/onq/2019/06/25/5g-rolling-out-globally-and-faster-any-g-it

[17] "Release description; Release 15," 3rd Generation Partnership Project (3GPP), Technical Report 21.915 V1.1.0, Mar 2019. [Online]. Available: https://portal.3gpp.org/desktopmodules/Specifications/SpecificationDetails.aspx?specificationId=3389

[18] "Release 16," 3rd Generation Partnership Project (3GPP), Jul 2019. [Online]. Available: https://www.3gpp.org/release-16

[19] "Release 17," 3rd Generation Partnership Project (3GPP), Sep 2019. [Online]. Available: https://www.3gpp.org/release-17

[20] "Guidelines for evaluation of radio interface technologies for IMT-2020," International Telecommunications Union (ITU), Report ITU-R M.2412-0, Oct 2017. [Online]. Available: https://www.itu.int/pub/R-REP-M.2412-2017

[21] A. Osseiran, F. Boccardi, V. Braun, K. Kusume, P. Marsch *et al.*, "Scenarios for 5G Mobile and Wireless Communications: the Vision of the METIS Project," *IEEE Communications Magazine*, vol. 52, no. 5, pp. 26–35, May 2014.

[22] C. Wang, F. Haider, X. Gao, X. You, Y. Yang *et al.*, "Cellular Architecture and Key Technologies for 5G Wireless Communication Networks," *IEEE Communications Magazine*, vol. 52, no. 2, pp. 122–130, Feb 2014.

[23] P. Popovski, J. J. Nielsen, C. Stefanovic, E. d. Carvalho, E. Strom *et al.*, "Wireless Access for Ultra-Reliable Low-Latency Communication: Principles and Building Blocks," *IEEE Network*, vol. 32, no. 2, pp. 16–23, Mar 2018.

[24] F. Schaich, B. Sayrac, M. Schubert, H. Lin, K. Pedersen *et al.*, "FANTASTIC-5G: 5G-PPP Project on 5G Air Interface Below 6 GHz," in *Proceedings of European Conference on Networks and Communications (EUCNC)*, Jun 2015. [Online]. Available: http://fantastic5g.com/wpcontent/uploads/2015/07/EuCNC-FANTASTIC-5G_final.pdf

[25] P. Popovski, K. F. Trillingsgaard, O. Simeone, and G. Durisi, "5G Wireless Network Slicing for eMBB, URLLC, and mMTC: A Communication-Theoretic View," *Computing Research Repository (CoRR)*, vol. abs/1804.05057, Aug 2018. [Online]. Available: http://arxiv.org/abs/1804.05057

[26] M. Bennis, M. Debbah, and H. V. Poor, "Ultrareliable and Low-Latency Wireless Communication: Tail, Risk, and Scale," *Proceedings of the IEEE*, vol. 106, no. 10, pp. 1834–1853, Oct 2018.

[27] "5G: A Technology Vision," Huawei Technologies Co. Ltd., Mar 2014. [Online]. Available: https://www.huawei.com/mediafiles/CORPORATE/PDF/Magazine/WinWin/HW_329327.pdf

[28] P. Popovski, "Ultra-reliable communication in 5G wireless systems," in *1st International Conference on 5G for Ubiquitous Connectivity*, Nov 2014, pp. 146–151.

[29] L. Wan, Z. Guo, Y. Wu, W. Bi, J. Yuan *et al.*, "4G/5G Spectrum Sharing: Efficient 5G Deployment to Serve Enhanced Mobile Broadband and Internet of Things Applications," *IEEE Vehicular Technology Magazine*, vol. 13, no. 4, pp. 28–39, Dec 2018.

[30] "Making 5G NR a reality," Qualcomm Technologies, Inc., Dec 2016. [Online]. Available: https://www.qualcomm.com/media/documents/files/whitepaper-making-5g-nr-a-reality.pdf

[31] "5G NR - A New Era for Enhanced Mobile Broadband," MediaTek, Mar 2018. [Online]. Available: https://cdn-www.mediatek.com/page/MediaTek-5G-NR-White-Paper-PDF5GNRWP.pdf

[32] "Study on physical layer enhancements for NR ultra-reliable and low latency case (URLLC)," 3rd Generation Partnership Project (3GPP), Technical Specification 38.824 V16.0.0, Mar 2019. [Online]. Available: https://portal.3gpp.org/desktopmodules/Specifications/SpecificationDetails.aspx?specificationId=3498

[33] M. R. Palattella, M. Dohler, A. Grieco, G. Rizzo, J. Torsner *et al.*, "Internet of Things in the 5G Era: Enablers, Architecture, and Business Models," *IEEE Journal on Selected Areas in Communications*, vol. 34, no. 3, pp. 510–527, Mar 2016.

[34] W. Yang, M. Wang, J. Zhang, J. Zou, M. Hua *et al.*, "Narrowband Wireless Access for Low-Power Massive Internet of Things: A Bandwidth Perspective," *IEEE Wireless Communications*, vol. 24, no. 3, pp. 138–145, Jun 2017.

[35] I. F. Akyildiz, S. Nie, S. Lin, and M. Chandrasekaran, "5G roadmap: 10 key enabling technologies"," *Computer Networks*, vol. 106, pp. 17–48, 2016. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1389128616301918

[36] M. Shirvanimoghaddam, M. Dohler, and S. J. Johnson, "Massive Non-Orthogonal Multiple Access for Cellular IoT: Potentials and Limitations," *IEEE Communications Magazine*, vol. 55, no. 9, pp. 55–61, Sep 2017.

[37] F. Ghavimi and H. Chen, "M2M Communications in 3GPP LTE/LTE-A Networks: Architectures, Service Requirements, Challenges, and Applications," *IEEE Communications Surveys Tutorials*, vol. 17, no. 2, pp. 525–549, Q2 2015.

[38] Z. Dawy, W. Saad, A. Ghosh, J. G. Andrews, and E. Yaacoub, "Toward Massive Machine Type Cellular Communications," *IEEE Wireless Communications*, vol. 24, no. 1, pp. 120–128, Feb 2017.

[39] S. Lien and K. Chen, "Massive Access Management for QoS Guarantees in 3GPP Machine-to-Machine Communications," *IEEE Communications Letters*, vol. 15, no. 3, pp. 311–313, 2011.

[40] A. Azari, "Energy-efficient scheduling and grouping for machine-type communications over cellular networks," *Ad Hoc Networks*, vol. 43, pp. 16–29, 2016. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1570870516300282

[41] K. Chatzikokolakis, A. Kaloxylos, P. Spapis, N. Alonistioti, C. Zhou *et al.*, "On the Way to Massive Access in 5G: Challenges and Solutions for Massive Machine Communications," in *International Conference on Cognitive Radio Oriented Wireless Networks*, M. Weichold, M. Hamdi, M. Shakir, M. Abdallah, G. Karagiannidis *et al.*, Eds. Springer, 2015, ch. 8, pp. 708–717.

[42] A. Zanella, N. Bui, A. Castellani, L. Vangelista, and M. Zorzi, "Internet of Things for Smart Cities," *IEEE Internet of Things Journal*, vol. 1, no. 1, pp. 22–32, Feb 2014.

[43] F. Hussain, A. Anpalagan, A. S. Khwaja, and M. Naeem, "Resource allocation and congestion control in clustered M2M communication using Q-learning," *Transactions on Emerging Telecommunications Technologies*, vol. 28, no. 4, pp. e3039–n/a, 2017.

[44] M. Koseoglu, "Lower Bounds on the LTE-A Average Random Access Delay Under Massive M2M Arrivals," *IEEE Transactions on Communications*, vol. 64, no. 5, pp. 2104–2115, May 2016.

[45] P. Zhang and G. Miao, "Energy-efficient clustering design for M2M communications," in *2014 IEEE Global Conference on Signal and Information Processing (GlobalSIP)*, Dec 2014, pp. 163–167.

[46] S. A. El-Feshawy, W. Saad, M. Shokair, and M. I. Dessouky, "An efficient clustering design for cellular based machine-to-machine communications," in *2018 35th National Radio Science Conference (NRSC)*, Mar 2018, pp. 177–186.

[47] U. Tefek and T. J. Lim, "Clustering and radio resource partitioning for machine-type communications in cellular networks," in *2016 IEEE Wireless Communications and Networking Conference*, Apr 2016, pp. 1–6.

[48] J. Chang, "A Distributed Cluster Computing Energy-Efficient Routing Scheme for Internet of Things Systems," *Wireless Personal Communications*, vol. 82, no. 2, pp. 757–776, May 2015.

[49] L. Dai, B. Wang, Y. Yuan, S. Han, C. I *et al.*, "Non-orthogonal multiple access for 5G: solutions, challenges, opportunities, and future research trends," *IEEE Communications Magazine*, vol. 53, no. 9, pp. 74–81, Sep 2015.

[50] Z. Yuan, G. Yu, W. Li, Y. Yuan, X. Wang *et al.*, "Multi-User Shared Access for Internet of Things," in *2016 IEEE 83rd Vehicular Technology Conference (VTC Spring)*, May 2016, pp. 1–5.

[51] M. Aldababsa, M. Toka, S. Gökçeli, G. K. Kurt, and O. Kucur, "A Tutorial on Nonorthogonal Multiple Access for 5G and Beyond,," *Wireless Communications and Mobile Computing*, Jun 2018. [Online]. Available: https://www.hindawi.com/journals/wcmc/2018/9713450/cta

[52] M. Vaezi, R. Schober, Z. Ding, and H. V. Poor, "Non-Orthogonal Multiple Access: Common Myths and Critical Questions," *Computing Research Repository (CoRR)*, vol. abs/1809.07224, 2018. [Online]. Available: http://arxiv.org/abs/1809.07224

[53] Z. Ding, M. Peng, and H. V. Poor, "Cooperative Non-Orthogonal Multiple Access in 5G Systems," *IEEE Communications Letters*, vol. 19, no. 8, pp. 1462–1465, Aug 2015.

[54] Z. Ding, P. Fan, and H. V. Poor, "Impact of User Pairing on 5G Nonorthogonal Multiple-Access Downlink Transmissions," *IEEE Transactions on Vehicular Technology*, vol. 65, no. 8, pp. 6010–6023, Aug 2016.

[55] Z. Ding, Y. Liu, J. Choi, Q. Sun, M. Elkashlan *et al.*, "Application of Non-Orthogonal Multiple Access in LTE and 5G Networks," *IEEE Communications Magazine*, vol. 55, no. 2, pp. 185–191, Feb 2017.

[56] X. Huang and N. Ansari, "Energy sharing within EH-enabled wireless communication networks," *IEEE Wireless Communications*, vol. 22, no. 3, pp. 144–149, Jun 2015.

[57] Y. Liu, Z. Ding, M. Elkashlan, and H. V. Poor, "Cooperative Non-orthogonal Multiple Access With Simultaneous Wireless Information and Power Transfer," *IEEE Journal on Selected Areas in Communications*, vol. 34, no. 4, pp. 938–953, Apr 2016.

[58] J. Gong and X. Chen, "Achievable Rate Region of Non-Orthogonal Multiple Access Systems With Wireless Powered Decoder," *IEEE Journal on Selected Areas in Communications*, vol. 35, no. 12, pp. 2846–2859, Dec 2017.

[59] P. D. Diamantoulakis, K. N. Pappi, Z. Ding, and G. K. Karagiannidis, "Wireless-Powered Communications With Non-Orthogonal Multiple Access," *IEEE Transactions on Wireless Communications*, vol. 15, no. 12, pp. 8422–8436, Dec 2016.

[60] Z. Yang, W. Xu, Y. Pan, C. Pan, and M. Chen, "Energy Efficient Resource Allocation in Machine-to-Machine Communications With Multiple Access and Energy Harvesting for IoT," *IEEE Internet of Things Journal*, vol. 5, no. 1, pp. 229–245, Feb 2018.

[61] M. A. Abd-Elmagid, A. Biason, T. ElBatt, K. G. Seddik, and M. Zorzi, "Non-Orthogonal Multiple Access schemes in Wireless Powered Communication Networks," in *2017 IEEE International Conference on Communications (ICC)*, May 2017, pp. 1–6.

[62] M. Sauter, *From GSM to LTE: An Introduction to Mobile Networks and Mobile Broadband*. John Wiley & Sons, Ltd, 2011.

[63] F. Gabriel, S. Wunderlich, S. Pandi, F. H. P. Fitzek, and M. Reisslein, "Caterpillar RLNC With Feedback (CRLNC-FB): Reducing Delay in Selective Repeat ARQ Through Coding," *IEEE Access*, vol. 6, pp. 44 787–44 802, Aug 2018.

[64] H. Ding, S. Ma, C. Xing, and Z. Fei, "Performance analysis of incremental redundancy hybrid ARQ in mobile ad hoc networks," in *2014 IEEE International Conference on Communications (ICC)*, Jun 2014, pp. 5759–5764.

[65] H. A. Ngo and L. Hanzo, "Hybrid Automatic-Repeat-reQuest Systems for Cooperative Wireless Communications," *IEEE Communications Surveys Tutorials*, vol. 16, no. 1, pp. 25–45, Q1 2014.

[66] G. C. Madueño, . Stefanović, and P. Popovski, "Efficient LTE access with collision resolution for massive M2M communications," in *2014 Globecom Workshops (GC Wkshps)*, Dec 2014, pp. 1433–1438.

[67] M. S. Ali, E. Hossain, and D. I. Kim, "LTE/LTE-A Random Access for Massive Machine-Type Communications in Smart Cities," *IEEE Communications Magazine*, vol. 55, no. 1, pp. 76–83, Jan 2017.

[68] J. Plachy, Z. Becvar, and E. C. Strinati, "Cross-layer approach enabling communication of high number of devices in 5G mobile networks," in *2015 IEEE 11th International Conference on Wireless and Mobile Computing, Networking and Communications (WiMob)*, 2015, pp. 809–816.

[69] M. Shirvanimoghaddam and S. J. Johnson, "Multiple Access Technologies for cellular M2M Communications: An Overview," *Computing Research Repository (CoRR)*, vol. abs/1611.05548, 2016. [Online]. Available: http://arxiv.org/abs/1611.05548

[70] R. C. Kizilirmak, "Non-Orthogonal Multiple Access (NOMA) for 5G Networks," in *Towards 5G Wireless Networks - A Physical Layer Perspective*, H. K. Bizaki, Ed. InTechOpen, 2016, ch. 4, pp. 83–98.

[71] Y. Saito, Y. Kishiyama, A. Benjebbour, T. Nakamura, A. Li *et al.*, "Non-Orthogonal Multiple Access (NOMA) for Cellular Future Radio Access," in *2013 IEEE 77th Vehicular Technology Conference (VTC Spring)*, Jun 2013, pp. 1–5.

[72] Z. Ding, X. Lei, G. K. Karagiannidis, R. Schober, J. Yuan *et al.*, "A Survey on Non-Orthogonal Multiple Access for 5G Networks: Research Challenges and Future Trends," *IEEE Journal on Selected Areas in Communications*, vol. 35, no. 10, pp. 2181–2195, Oct 2017.

[73] K. Higuchi and A. Benjebbour, "Non-orthogonal Multiple Access (NOMA) with Successive Interference Cancellation for Future Radio Access," *IEICE Transactions on Communications*, vol. E98.B, no. 3, pp. 403–414, 2015.

[74] S. M. R. Islam, M. Zeng, and O. A. Dobre, "NOMA in 5G Systems: Exciting Possibilities for Enhancing Spectral Efficiency," *Computing Research Repository (CoRR)*, vol. abs/1706.08215, 2017. [Online]. Available: http://arxiv.org/abs/1706.08215

[75] H. Tabassum, M. S. Ali, E. Hossain, M. J. Hossain, and D. I. Kim, "Uplink Vs. Downlink NOMA in Cellular Networks: Challenges and Research Directions," in *2017 IEEE 85th Vehicular Technology Conference (VTC Spring)*, Jun 2017, pp. 1–7.

[76] M. S. Ali, H. Tabassum, and E. Hossain, "Dynamic User Clustering and Power Allocation for Uplink and Downlink Non-Orthogonal Multiple Access (NOMA) Systems," *IEEE Access*, vol. 4, pp. 6325–6343, Aug 2016.

[77] S. M. R. Islam, N. Avazov, O. A. Dobre, and K. Kwak, "Power-Domain Non-Orthogonal Multiple Access (NOMA) in 5G Systems: Potentials and Challenges," *IEEE Communications Surveys Tutorials*, vol. 19, no. 2, pp. 721–742, Q2 2017.

[78] M. R. Usman, A. Khan, M. A. Usman, Y. S. Jang, and S. Y. Shin, "On the performance of perfect and imperfect SIC in downlink non orthogonal multiple access (NOMA)," in *2016 International Conference on Smart Green Technology in Electrical and Information Systems (ICSGTEIS)*, Oct 2016, pp. 102–106.

[79] D. Cai, Z. Ding, P. Fan, and Z. Yang, "On the Performance of NOMA With Hybrid ARQ," *IEEE Transactions on Vehicular Technology*, vol. 67, no. 10, p. 10033–10038, Oct 2018.

[80] "Nomenclature of the frequency and wavelength bands used in telecommunications," International Telecommunications Union (ITU), Recommendation ITU-R V.431-8, Aug 2015. [Online]. Available: https://www.itu.int/rec/R-REC-V.431-8-201508-I

[81] "Framework and overall objectives of the future development of IMT for 2020 and beyond," International Telecommunications Union (ITU), Recommendation ITU-R M.2083-0, Sep 2015. [Online]. Available: https://www.itu.int/rec/R-REC-M.2083-0-201509-I

[82] "Designing for the future: the 5G NR physical layer," Ericsson, Jul 2017. [Online]. Available: https://www.ericsson.com/en/ericsson-technology-review/archive/2017/designing-for-the-future-the-5g-nr-physical-layer

[83] A. A. Zaidi, R. Baldemair, V. Moles-Cases, N. He, K. Werner *et al.*, "OFDM Numerology Design for 5G New Radio to Support IoT, eMBB, and MBSFN," *IEEE Communications Standards Magazine*, vol. 2, no. 2, pp. 78–83, Jun 2018.

[84] K. Takeda, L. H. Wang, and S. Nagata, "Industry Perspectives Latency Reduction Toward 5G," *IEEE Wireless Communications*, vol. 24, no. 3, pp. 2–4, Jun 2017.

[85] K. Zhou and N. Nikaein, "Packet aggregation for machine type communications in LTE with random access channel," in *2013 IEEE Wireless Communications and Networking Conference (WCNC)*, Apr 2013, pp. 262–267.

[86] A. Laya, L. Alonso, and J. Alonso-Zarate, "Is the Random Access Channel of LTE and LTE-A Suitable for M2M Communications? A Survey of Alternatives," *IEEE Communications Surveys Tutorials*, vol. 16, no. 1, pp. 4–16, Q1 2014.

[87] D. Feng, C. She, K. Ying, L. Lai, Z. Hou *et al.*, "Towards Ultra-Reliable Low-Latency Communications: Typical Scenarios, Possible Solutions, and Open Issues," *IEEE Vehicular Technology Magazine*, vol. 14, no. 2, pp. 94–102, Apr 2019.

[88] P. Schulz, M. Matthe, H. Klessig, M. Simsek, G. Fettweis *et al.*, "Latency Critical IoT Applications in 5G: Perspective on the Design of Radio Interface and Network Architecture," *IEEE Communications Magazine*, vol. 55, no. 2, pp. 70–78, Feb 2017.

[89] C. Wang, Y. Chen, Y. Wu, and L. Zhang, "Performance Evaluation of Grant-Free Transmission for Uplink URLLC Services," in *2017 IEEE 85th Vehicular Technology Conference (VTC Spring)*, Jun 2017, pp. 1–6.

[90] T. Fehrenbach, R. Datta, B. Göktepe, T. Wirth, and C. Hellge, "URLLC Services in 5G Low Latency Enhancements for LTE," in *2018 IEEE 88th Vehicular Technology Conference (VTC-Fall)*, Aug 2018, pp. 1–6.

[91] P. Si, J. Yang, S. Chen, and H. Xi, "Adaptive Massive Access Management for QoS Guarantees in M2M Communications," *IEEE Transactions on Vehicular Technology*, vol. 64, no. 7, pp. 3152–3166, Jul 2015.

[92] S. Lien, Y. Lin, and K. Chen, "Cognitive and Game-Theoretical Radio Resource Management for Autonomous Femtocells with QoS Guarantees," *IEEE Transactions on Wireless Communications*, vol. 10, no. 7, pp. 2196–2206, Jul 2011.

[93] A. G. Gotsis, A. S. Lioumpas, and A. Alexiou, "Evolution of packet scheduling for Machine-Type communications over LTE: Algorithmic design and performance analysis," in *2012 IEEE Globecom Workshops*, Dec 2012, pp. 1620–1625.

[94] A. Azari and G. Miao, "Energy efficient MAC for cellular-based M2M communications," in *2014 IEEE Global Conference on Signal and Information Processing (GlobalSIP)*, 2014, pp. 128–132.

[95] A. H. Mohammed, A. S. Khwaja, A. Anpalagan, and I. Woungang, "Base Station Selection in M2M Communication Using Q-Learning Algorithm in LTE-A Networks," in *2015 IEEE 29th International Conference on Advanced Information Networking and Applications*, Mar 2015, pp. 17–22.

[96] P. Schulz, A. Wolf, G. P. Fettweis, A. M. Waswa, D. Mohammad Soleymani *et al.*, "Network Architectures for Demanding 5G Performance Requirements: Tailored Toward Specific Needs of Efficiency and Flexibility," *IEEE Vehicular Technology Magazine*, vol. 14, no. 2, pp. 33–43, Jun 2019.

[97] P. Tan, M. Steinbach, and V. Kumar, "Cluster Analysis: Basic Concepts and Algorithms," in *Introduction to Data Mining*, 1st ed. New York, USA: Pearson Education, Inc., 2006, ch. 8, pp. 487–568.

[98] P. Sasikumar and S. Khara, "K-Means Clustering in Wireless Sensor Networks," in *2012 Fourth International Conference on Computational Intelligence and Communication Networks*, Nov 2012, pp. 140–144.

[99] W. Fakhet, S. E. Khediri, A. Dallali, and A. Kachouri, "New K-means algorithm for clustering in wireless sensor networks," in *2017 International Conference on Internet of Things, Embedded Systems and Communications (IINTEC)*, Oct 2017, pp. 67–71.

[100] G. Kumar, H. Mehra, A. R. Seth, P. Radhakrishnan, N. Hemavathi *et al.*, "An hybrid clustering algorithm for optimal clusters in Wireless sensor networks," in *2014 IEEE Students' Conference on Electrical, Electronics and Computer Science (SCEECS)*, Mar 2014, pp. 1–6.

[101] D. Tse and P. Viswanath, *Fundamentals of Wireless Communication*. New York, NY, USA: Cambridge University Press, 2005.

[102] "Study on Downlink Multiuser Superposition Transmission (MUST) for LTE (Release 13)," 3rd Generation Partnership Project (3GPP), Technical Specification 36.859 V13.0.0, Dec 2015. [Online]. Available: https://portal.3gpp.org/desktopmodules/Specifications/SpecificationDetails.aspx?specificationId=2912

[103] R. Jain, D. Chiu, and W. Hawe, "A Quantitative Measure Of Fairness And Discrimination For Resource Allocation In Shared Computer Systems," *Computing Research Repository (CoRR)*, vol. cs.NI/9809099, 1998. [Online]. Available: http://arxiv.org/abs/cs.NI/9809099

[104] M. Al-Imari, P. Xiao, M. A. Imran, and R. Tafazolli, "Uplink non-orthogonal multiple access for 5G wireless networks," in *2014 11th International Symposium on Wireless Communications Systems (ISWCS)*, Aug 2014, pp. 781–785.

[105] Z. Wei, J. Guo, D. W. K. Ng, and J. Yuan, "Fairness Comparison of Uplink NOMA and OMA," in *2017 IEEE 85th Vehicular Technology Conference (VTC Spring)*, Jun 2017, pp. 1–6.

[106] S. Timotheou and I. Krikidis, "Fairness for Non-Orthogonal Multiple Access in 5G Systems," *IEEE Signal Processing Letters*, vol. 22, no. 10, pp. 1647–1651, Oct 2015.

[107] T. N. Do, D. B. da Costa, T. Q. Duong, and B. An, "Improving the Performance of Cell-Edge Users in NOMA Systems Using Cooperative Relaying," *IEEE Transactions on Communications*, vol. 66, no. 5, pp. 1883–1901, May 2018.

[108] X. Zhang, W. Cheng, and H. Zhang, "Full-duplex transmission in phy and mac layers for 5G mobile wireless networks," *IEEE Wireless Communications*, vol. 22, no. 5, pp. 112–121, Oct 2015.

[109] M. F. Kader, S. Y. Shin, and V. C. M. Leung, "Full-Duplex Non-Orthogonal Multiple Access in Cooperative Relay Sharing for 5G Systems," *IEEE Transactions on Vehicular Technology*, vol. 67, no. 7, pp. 5831–5840, Jul 2018.

[110] Z. Zhang, Y. Liu, F. Luo, and X. Wang, "On Cancellation Capability of Full-Duplex RF Self-Interference Cancellation Schemes," *Mobile Information Systems*, Mar 2019. [Online]. Available: https://www.hindawi.com/journals/misy/2019/5627178/cta/

[111] H. Alves, R. D. Souza, and M. E. Pellenz, "Brief survey on full-duplex relaying and its applications on 5G," in *2015 IEEE 20th International Workshop on Computer Aided Modelling and Design of Communication Links and Networks (CAMAD)*, Sep. 2015, pp. 17–21.

[112] A. Nadh, R. Jagannath, and R. K. Ganti, "Self-interference cancellation in full-duplex wireless devices: A survey," *CSI Transactions on ICT*, vol. 7, no. 1, pp. 3–12, Mar 2019.

[113] R. Sattiraju and H. D. Schotten, "Reliability Modeling, Analysis and Prediction of Wireless Mobile Communications," in *2014 IEEE 79th Vehicular Technology Conference (VTC Spring)*, May 2014, pp. 1–6.

[114] D. Arthur and S. Vassilvitskii, "K-means++: The Advantages of Careful Seeding," in *Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms*, ser. SODA '07, 2007, pp. 1027–1035. [Online]. Available: http://dl.acm.org/citation.cfm?id=1283383.1283494

[115] B. Bahmani, B. Moseley, A. Vattani, R. Kumar, and S. Vassilvitskii, "Scalable K-Means++," *Proc. VLDB Endow.*, vol. 5, no. 7, pp. 622–633, Mar 2012.

[116] L. Kaufman and P. J. Rousseeuw, *Finding Groups in Data: An Introduction to Cluster Analysis*. New York, USA: John Wiley & Sons, 1990.

[117] "Evolved Universal Terrestrial Radio Access (E-UTRA); Further advancements for E-UTRA physical layer aspects," 3rd Generation Partnership Project (3GPP), Technical Specification 36.814 V9.2.0, Mar 2017. [Online]. Available: https://portal.3gpp.org/desktopmodules/Specifications/SpecificationDetails.aspx?specificationId=2493

[118] M. Woltering, D. Wubben, A. Dekorsy, V. Braun, and U. Doetsch, "Link Level Performance Assessment of Reliability-Based HARQ Schemes in LTE," in *2014 IEEE 79th Vehicular Technology Conference (VTC Spring)*, May 2014, pp. 1–5.

[119] R. D. Wesel, N. Wong, A. Baldauf, A. Belhouchat, A. Heidarzadeh *et al.*, "Transmission Lengths That Maximize Throughput of Variable-Length Coding & ACK-/NACK Feedback," in *2018 IEEE Global Communications Conference (GLOBECOM)*, Dec 2018, pp. 1–6.

[120] J. Choi, "H-ARQ Based Non-Orthogonal Multiple Access with Successive Interference Cancellation," in *IEEE GLOBECOM 2008 - 2008 IEEE Global Telecommunications Conference*, Nov 2008, pp. 1–5.

[121] A. Li, A. Benjebbour, X. Chen, H. Jiang, and H. Kayama, "Investigation on hybrid automatic repeat request (HARQ) design for NOMA with SU-MIMO," in *2015 IEEE 26th Annual International Symposium on Personal, Indoor, and Mobile Radio Communications (PIMRC)*, Aug 2015, pp. 590–594.

[122] X. Li, "An Enhanced Hybrid Automatic Repeat Request (HARQ) Scheme for Non-orthogonal Multiple Access (NOMA) System of 5G," in *2016 International Conference on Information Engineering and Communications Technology (IECT 2016)*, Nov 2016, pp. 1–5.

[123] J. Choi, "On HARQ-IR for Downlink NOMA Systems," *IEEE Transactions on Communications*, vol. 64, no. 8, pp. 3576–3584, Aug 2016.

[124] D. Cai, Y. Xu, F. Fang, Z. Ding, and P. Fan, "On the Impact of Time-Correlated Fading for Downlink NOMA," *IEEE Transactions on Communications*, vol. 67, no. 6, pp. 4491–4504, Jun 2019.

[125] R. Kotaba, C. N. Manchon, N. M. K. Pratas, T. Balercia, and P. Popovski, "Improving Spectral Efficiency in URLLC via NOMA-Based Retransmissions," in *ICC 2019 - 2019 IEEE International Conference on Communications (ICC)*, May 2019, pp. 1–7.

[126] J. Östman, R. Devassy, G. C. Ferrante, and G. Durisi, "Low-Latency Short-Packet Transmissions: Fixed Length or HARQ?" *Computing Research Repository (CoRR)*, vol. abs/1809.06560, 2018. [Online]. Available: http://arxiv.org/abs/1809.06560

[127] G. Caire and D. Tuninetti, "The throughput of hybrid-ARQ protocols for the Gaussian collision channel," *IEEE Transactions on Information Theory*, vol. 47, no. 5, pp. 1971–1988, Jul 2001.

[128] Z. Ding and H. V. Poor, "Design of Massive-MIMO-NOMA With Limited Feedback," *IEEE Signal Processing Letters*, vol. 23, no. 5, pp. 629–633, May 2016.

[129] Z. Yu, C. Zhai, and J. Liu, "Non-orthogonal multiple access relaying with truncated ARQ," *IET Communications*, vol. 11, no. 4, pp. 514–521, 2017.

[130] Z. Yu, C. Zhai, W. Ni, and D. Wang, "Non-Orthogonal Multiple Access With Cooperative Truncated ARQ and Relay Selection," *IEEE Access*, vol. 7, pp. 56 228–56 243, Apr 2019.

[131] J. Kim and I. Lee, "Non-Orthogonal Multiple Access in Coordinated Direct and Relay Transmission," *IEEE Communications Letters*, vol. 19, no. 11, pp. 2037–2040, Nov 2015.

[132] Y. Tao, L. Liu, S. Liu, and Z. Zhang, "A survey: Several technologies of non-orthogonal transmission for 5G," *China Communications*, vol. 12, no. 10, pp. 1–15, Oct 2015.

[133] D. Zhang, M. Matthé, L. L. Mendes, and G. Fettweis, "A Study on the Link Level Performance of Advanced Multicarrier Waveforms Under MIMO Wireless Communication Channels," *IEEE Transactions on Wireless Communications*, vol. 16, no. 4, pp. 2350–2365, Apr 2017.

[134] G. Wunder, P. Jung, M. Kasparick, T. Wild, F. Schaich *et al.*, "5GNOW: non-orthogonal, asynchronous waveforms for future mobile applications," *IEEE Communications Magazine*, vol. 52, no. 2, pp. 97–105, Feb 2014.

[135] X. He, F. Wang, X. Chen, D. Miao, and Z. Zhao, "Non-orthogonal waveforms for machine type communication," in *2017 XXXIInd General Assembly and Scientific Symposium of the International Union of Radio Science (URSI GASS)*, Aug 2017, pp. 1–4.

[136] J. Kim, Y. Park, S. Weon, J. Jeong, S. Choi *et al.*, "A New Filter-Bank Multicarrier System: The Linearly Processed FBMC System," *IEEE Transactions on Wireless Communications*, vol. 17, no. 7, pp. 4888–4898, Jul 2018.

[137] Z. Ding, L. Dai, and H. V. Poor, "MIMO-NOMA Design for Small Packet Transmission in the Internet of Things," *IEEE Access*, vol. 4, pp. 1393–1405, Apr 2016.

[138] L. Dai, B. Wang, M. Peng, and S. Chen, "Hybrid Precoding-Based Millimeter-Wave Massive MIMO-NOMA With Simultaneous Wireless Information and Power Transfer," *IEEE Journal on Selected Areas in Communications*, vol. 37, no. 1, pp. 131–141, Jan 2019.

[139] Y. Ai and M. Cheffena, "Performance Analysis of Hybrid-ARQ over Full-Duplex Relaying Network Subject to Loop Interference under Nakagami-m Fading Channels," in *2017 IEEE 85th Vehicular Technology Conference (VTC Spring)*, Jun 2017, pp. 1–5.

[140] D. Wang, R. Zhang, X. Cheng, and L. Yang, "Capacity-Enhancing Full-Duplex Relay Networks based on Power-Splitting (PS-)SWIPT," *IEEE Transactions on Vehicular Technology*, vol. 66, no. 6, pp. 5445–5450, Jun 2017.