RECEIVER ARCHITECTURES AND ALGORITHMS FOR NON-ORTHOGONAL MULTIPLE ACCESS

Talgat Manglayev

A thesis submitted in partial fulfillment of the requirement of Nazarbayev University for the degree of Doctor of Philosophy

May 2020

Abstract

Multiple access (MA) schemes in cellular systems aim to provide high throughput to multiple users simultaneously while utilising the network resources efficiently. Traditionally, each user in the network is assigned a fraction of resources (such as slots in time or frequency) to operate so that multi-user interference is avoided. These schemes are named as 'orthogonal multiple access' (OMA) and are the basis of most cellular standards – from the earliest first generation up to the current fourth-generation systems. Non-orthogonal multiple access (NOMA) on the other hand is a novel method that allows all the users in the network to operate in the entire available spectrum at the same time which enables significant improvement in the system throughput.

While providing increased throughput, NOMA requires high computational power in order to implement sophisticated interference cancellation algorithms at each user terminal, as well as power allocation schemes at the base station. As a potential candidate for the fifth-generation networks (5G), NOMA must meet certain requirements, and computational efficiency is essential for reduced latency. Recently graphics processing units (GPUs), which were initially intended for outputting images to display, appeared as an alternative to multi-core central processing units (CPUs) for general-purpose computing. GPUs have thousands of cores with approximately three times less frequency than a CPU core. With their numerous advantages in executing heavy and time-consuming computations in parallel, GPUs have become attractive platforms in a variety of fields.

The overall aim of this research is to significantly increase the scientific understanding and technical knowledge on NOMA. This is achieved by exploring and developing novel methods, models, designs and techniques that will facilitate the implementation of NOMA for future generation networks.

First, the achievable data rates for individual users are demonstrated in a successful interference cancellation (SIC) based NOMA network. These results were compared against the conventional orthogonal MA schemes with optimum power allocation and varying fairness. In addition, a further investigation was carried out into the deficiency of SIC receivers which can occur when a user in the networks attempts to decode other users' signal. Presented in the analysis is the findings from the experimental process where the decoding order of a user with a mismatched signal was observed as well as the significant impact on the computation time. The decoding time-difference between correct and mismatched decoding order as a detection method of deficiency or fraudulence in the network is then discussed. Next, a comparison is presented between the computational times of the SIC receiver with another popular interference cancellation scheme named 'parallel interference cancellation' (PIC). This was done using different platforms specifically for an uplink NOMA system. The results showed that the computation time of PIC scheme is significantly lower than SIC on the GPU platform even for a very large number of available users in the network. Then, the execution time of NOMA with SIC in the uplink of a cellular network with user clustering was examined. User clustering is a popular method in NOMA networks that eases the sophisticated resource allocation and network management issues. While most works found in the literature review concentrate on the joint optimisation of user grouping and resources, this research project focused on processing the signal detection of each cluster in parallel on the GPU platform at the base station. Following this, parallel interference cancellation (PIC) was implemented and compared with the existing SIC on both CPU and GPU platforms for uplink NOMA-OFDM. Architectures of the receivers were modified to fit into parallel processing. GPU was found applicable to speed up computations in NOMA based next-generation cellular networks outperforming up to 220 times SIC on CPU. Finally, the research presents the power allocation problem from artificial intelligence (AI) perspective and propose a method to predict the power allocation coefficients in a downlink NOMA system. The results of the research show a close-to-optimal sum rate with about 120 times reduced

computation time. The achieved results decreases the network latency and assist NOMA to meet 5G requirements.

Contents

Chapter 1

Introduction

This chapter deals with a history of wireless cellular networks (Section 1.1); potential techniques of the future radio access (Section 1.2); history of recent computation (Section 1.3); novelty, contribution, motivation, aims and objectives of the work (Section 1.4) and finally the thesis organisation (Section 1.6). Specifically, the history section carries background information of the thesis and narrows down to the topic focus. Previously used and potential techniques are given as background information. Similarly, we discussed progress and varieties of computing processors then we describe the main idea of the thesis by presenting novelty, contribution, hypothesis, aim and objectives of the work. Finally, the thesis organisation is given.

1.1 A Brief History of Wireless Cellular Networks

Wireless communications as a branch of information and communication technologies have undergone a rapid development tendency. According to Guturu [1], a new generation of cellular networks traditionally launches at the beginning of each decade since 1980s. First generation (1G) which introduced standard voice calls started in 1982. The voice calls used the frequency division multiple access (FDMA) technique in the physical layer. They were digitally encrypted in the second-generation (2G) a decade later. Digital signals introduced new services such as sending multimedia and short text messages.

Moreover, 2G achieved spectral efficiency of available bandwidth. The 2G standards implemented the time division multiple access (TDMA) scheme in Europe (as in GSM) and the code division multiple access (CDMA) scheme for North America (as in cdmaOne). New possibilities such as mobile TV, data services and video calls appeared later in the third generation (3G) networks. The 3G Partnership Project (3GPP) and the International Mobile Telecommunications (IMT2000) launched standardisation towards 3G in 1998. The latter required 200 kbits/s data rates as standard to be called '3G'. The Universal Mobile Telephone System (UMTS) and the CDMA 2000 became the 3G standards. They began in Europe, Japan and China in 2001 and were later launched in North America and Korea in 2002. Further improvements in 3G resulted in the release of 3.5G and 3.75G which made broadband internet available on a massive scale by providing data rates of up to 14.7 Mb/s for smartphones and mobile modems. Fourth generation (4G) started in 2009 with two systems: mobile World Interoperability for Microwave Access (WiMAX) based on IEEE 802.16 standards and long-term evolution (LTE) standards. Mobile services were extended with internet access, internet protocol telephony, high-definition mobile TV, video conferencing and video games in 4G. High data rates attracted a large number of connected devices and the appearance of high bandwidth applications.

At the moment both the industry and academia are engaged in researching the requirements and solutions for the next generation (5G). The main technical challenges are: to achieve a data rate between 1 to 10 Gb/sec, reduce latency to less than 1 millisecond (ms), extend bandwidth for a large number of devices with full availability and coverage and finally reduce energy consumption. These goals for 5G are based on insights of telecommunication companies such as Nokia, Huawei and Docomo. The enterprises are aiming for massive connection of devices, internal communications, ultra-high-speed and deployment of the efficient spectrum on top of enhancement in all existing network features. The Internet of Things (IoT) and Machine-to-Machine (M2M) communications are examples of such massive connectivity and offer a variety of network services. For example, ultra-high-speed throughput will make possible virtual reality, high definition video

calls and 4K video broadcasting. Similarly, the deployment of the efficient spectrum will enable the launch of demanding applications and a large number of equipment per user with a variety of complicated scenarios. Such predictions are based on the evolution of previous generations and related technologies e.g. mobile hardware, applications etc.

Wireless communication systems are going to change ordinary daily life in various areas including health, industry and logistics [2–4]. Complex cellular network architecture and design with novel technologies cover every layer and promise to reach the set objectives. Some of the on-going and expected advancements for cellular networks are briefly mentioned in this paragraph. One of the simulations and tests of 5G with commercial instruments allows recognition of data flow and configuration of the cell with different parameters [5]. Cell design involves advanced technology like centralised radio access network (C-RAN) [6] and multiple-input multiple-output (MIMO) [7]. As well as the graphics processing unit (GPU) cluster is offered [8] to outsource complex gaming computations for mobile devices.

Radio frequency selection is a challenge for industry and scholars whereas its allocation remains an issue for the governments. On the one hand frequency bands higher than 29 GHz require deployments of lots of base stations, whereas lower frequency bands may be already reserved for military, rescue or other needs. For example, the expected frequency bands for Europe in the 5G as mentioned in [9] were below 5 GHz. Massive machine communication expected in 5G has another challenge in realisation. Preliminary requirements allow from 10 to 100 times more connected devices and long-life battery for more than 5 years with lower costs and 99% coverage. However, the issue is in the number of machines that are not widespread enough for such needs [10]. The multiple access scheme is established as a distinguishing physical layer feature of any wireless communication network. Orthogonal multiple access techniques have been implemented in wireless cellular networks from the beginning and up to the current fourth generation [11].

1.2 Potential Techniques of Future Radio Access

One of the proposed schemes to meet 5G requirements discussed above provides multiple connections either in-power or in-code domain and remains non-orthogonal in time and frequency [12]. Examples of non-orthogonal multiple access (NOMA) schemes in the power domain are based on power level differences of signals of the connected devices. On the other hand, NOMA in code domain assigns a unique code to the signal which is used to demodulate the message at the receiver side. Since non-orthogonality introduces interference among the signals the receiver requires an interference cancellation technique. It must be reliable, practical and most of all, it must meet 5G requirements.

Successive interference cancellation (SIC) and parallel interference cancellation (PIC) techniques are offered with NOMA within the power domain in the literature. The SIC decodes the signals of each user sequentially. It starts by decoding the signal that belongs to the user with the highest power, then subtracts it from the received signal, and then it iterates the same process for the second-highest power signal. In PIC, which is an alternative way, the signals of all users are decoded concurrently [13], [14]. NOMA's performance promises to meet the 5G requirements which is what made it 'trendy' among candidate technologies, however, feasibility challenges are preventing the deployment of the scheme. One of them is controlling the cellular network during the different level of traffic load. Also, the reliability requires correct interference cancellation, for example, in every iteration in SIC receiver. The work in [15] indicated key obstacles of NOMA deployment in 5G. Firstly, the authors argue that the power domain is still on an infancy research level rather than the practical one. Secondly, the signal processing technology of the chip at the receiver side tends to require design improvements as SIC computation is still complex.

1.3 The Recent Progress of Computation

Computation cost has been a challenge for both software and hardware engineers. Software engineers keep reducing the number of instructions and search for creative algorithms to reduce the running time of a program, whereas hardware engineers devise solutions that save physical resources. Blelloch in [16] describes a parallel programming concept and programming language named “the NESL" which allows parallel operations that reduce the number of operations. Massive pioneer hardware solutions of parallel programming appeared a decade later than the idea. ® Core ™ Duo processors introduced by Intel Inc. in 2006 offered architecture with two cores and shared cache memory. The challenges involved virtualisation, power and temperature control as described in [17]. The latest huge core i9 processors of the 9th generation by Intel Inc. have up to 18 physical cores with 3.0 GHz frequency [18]. Multi-threaded programs are needed in order to use those multiple cores to realise tangible performance benefits. There are various common programming languages such as assembler, C, Java, C# etc., where multi-threading is possible [19]. If the number of initiated threads exceeds the number of cores then threads will not be executed simultaneously.

The processing speed of a graphics processing units (GPU) core is usually approximately two to three times less than the frequency of a CPU core. For example, one of the latest models of NVIDIA GPU RTX 2080 has a 1515 MHz clock speed. In contrast, the number of cores in GPU is counted in hundreds, so the same RTX 2080 has 2944 CUDA cores against 18 cores in the latest produced CPU as mentioned above. Hardware characteristics of GPU allow general-purpose computing and high-performance computing (HPC) using the advantage of numerous multiple cores upon parallel algorithms.

Indeed, digital signal processors (DSP) are naturally suitable devices for encoding and decoding. Moreover, they are flexible and suit to a number of operating systems. The processors may be built-in to both the trendy IoT devices and to the BS. Hence, they fit both downlink and uplink channels. However, the classic DSPs are obsolete and are not able compete with modern GPUs with a major parameter like GFLOPS [20]. The DSP

enthusiasts lost the computation competition to GPU because of the ability to run tasks in parallel of the latter one [21].

In [22] first HPC cluster with 30 GPUs for scientific computation was assembled and demonstrated. Some of the most popular platforms which enable general-purpose computing and HPC are OpenCL from Intel®, CUDA from NVIDIA® and MATLAB® parallel computing toolbox from Mathworks. This leads to a speedup of heavy and time-consuming research simulations and computations in different areas. Different ways of accelerating NOMA related techniques with parallel programming using multiple CPU and GPU threads are described in this thesis.

1.4 Aims and Objectives

This thesis aims to increase the scientific understanding and the technical knowledge on non-orthogonal multiple access (NOMA) systems by exploring and developing novel methods, models, designs and techniques that will facilitate the implementation of future radio access. Towards this aim the following objectives are pursued:

1. To develop a novel interference cancellation mechanism for improved capacity.
2. To involve GPU for multi-user decoding in NOMA networks.
3. To investigate the feasibility of parallel programming approaches for multi-user decoding.

1.5 List of Publications

Journal Publications

[1] T. Manglayev, R. C. Kizilirmak, N.A.W.A. Hamid “GPU Accelerated PIC and SIC for OFDM-NOMA" *Electronics*, vol. 8, no. 3, pp. Feb. 2019.

[2] T. Manglayev, R. C. Kizilirmak, Y.H. Kho, N.A.W.A. Hamid “GPU Accelerated Successive Interference Cancellation for NOMA Uplink with User Clustering", *Wireless Personal Communications*, vol. 103 no. 3, pp. 2391-2400, Dec. 2018.

Conference Publications

[1] T. Manglayev, R.C. Kizilirmak, Y.H. Kho “Comparison Of Parallel And Successive Interference Cancellation For Non-Orthogonal Multiple Access" *Comparison Of Parallel And Successive Interference Cancellation For Non-Orthogonal Multiple Access*, Astana, 2018 pp. 74-77.

[2] T. Manglayev, R.C. Kizilirmak, Y.H. Kho, N. Bazhayev, I. Lebedev “NOMA with imperfect SIC implementation", *IEEE EUROCON 2017 -17th International Conference on Smart Technologies*, 2017, Ohrid, pp. 22-25.

[3] T. Manglayev, R.C. Kizilirmak, Y.H. Kho “Optimum power allocation for non-orthogonal multiple access (NOMA)", *2016 IEEE 10th International Conference on Application of Information and Communication Technologies (AICT)*, Baku, 2016, pp. 1-4.

1.6 Thesis Organization

There are six chapters in the thesis. The first chapter is the introduction to the thesis. It discusses the historical milestones that led to the current stage of wireless communications and discusses a forward view of future radio access. The chapter also states the aim and objectives of the research and describes the contributions of the work.

The next chapter presents preliminary and fundamental work which is relevant to the research. Firstly, works related to multiple access schemes are reviewed followed by a review of NOMA. After discussing the basics of NOMA in both uplink and downlink channels, the two most popular interference cancellation techniques were introduced, namely

successive interference cancellation (SIC) and parallel interference cancellation (PIC). Next, power allocation and its vital role in power domain NOMA are discussed followed by user grouping in the cellular systems. Then, we switch our focus to parallel programming concept. Examples of advanced engineering solutions are offered where parallel programming is applied to wireless communications. Finally, the recent role of artificial intelligence (AI) in the upcoming mobile systems are presented with a particular focus on 5G.

Chapter three has three subsections. The first topic is optimum power allocation for NOMA networks with SIC receiver. Power allocation coefficients are obtained to reach the maximum sum rate for the users with a predefined fairness constraint. The sum-rate results of NOMA and OMA are compared and the superiority of NOMA is demonstrated. The second subsection discusses the potential threat in the receivers and discusses a method that identifies it. Decoding time for the users in a network is proportional to their distance to the base station. The possible attack can be detected by tracking the decoding times of the users in the network. The third part studies NOMA with SIC and PIC receivers in the uplink channel. These two receivers are the most common in the literature and their decoding times will be measured and compared using different platforms: MATLAB using ordinary CPU; Java programming language that runs multithreaded CPU and CUDA platform with GPU.

Chapter four proposes solutions for enhancing the computation time of the NOMA receiver for reduced latency. The first part considers user clustering for NOMA with SIC in the uplink channel. The second part focuses on OFDM-NOMA with SIC and PIC receivers. Both parts present the receivers implemented on a GPU device as an alternative to CPU and present the results for comparison.

Chapter five demonstrates machine learning (ML) and deep learning (DL) from the AI family to assist the power allocation mechanism in NOMA scheme. Numerical results evidence that AI-enabled power allocation gives very close to optimum power allocation in terms of sum capacity and the execution times of ML and DL is much faster than

exhaustive search.

Chapter six concludes the thesis and discusses potential future works that may further develop the idea. The long-term and short-term perspectives are presented separately.

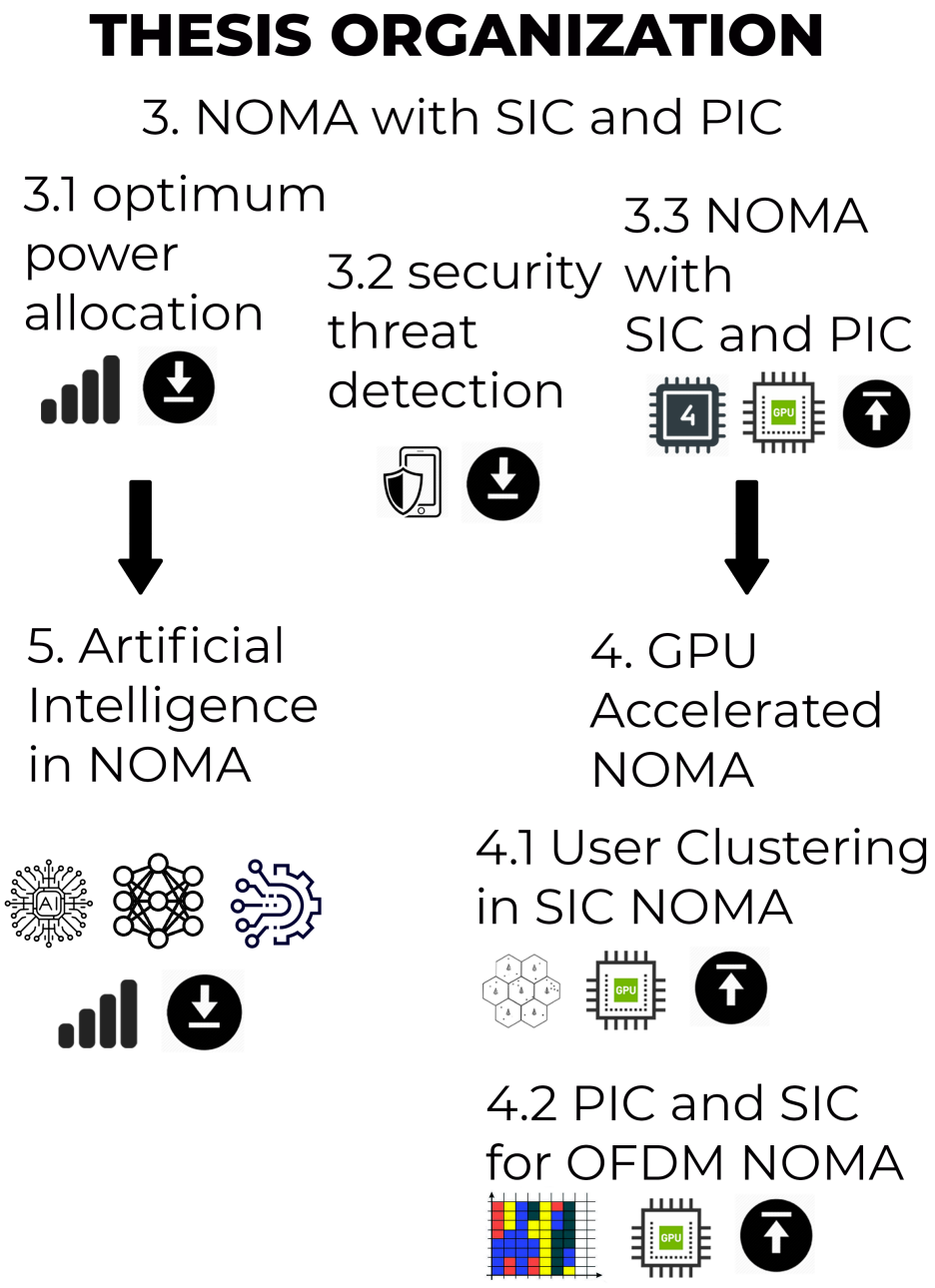

FIGURE 1.1: Info graphics of the main parts of the thesis.

Chapter 2

Background and Preliminaries

This chapter deals with the research and findings relating to multiple access schemes in general (Section 2.1), NOMA in power domain (Section 2.2), interference cancellation in NOMA (Section 2.3), NOMA user clustering (Section 2.4) and parallel programming (Section 2.5). Specifically, multiple access schemes in earlier cellular networks and variants of NOMA which have the potential to be implemented into 5G are reviewed. The SIC and PIC decoding schemes at the NOMA receiver are also discussed. Following this, literature about user-clustering in NOMA that is grouping the users in a cell for improved performance is then reviewed. Finally, the focus is switched to the computation part of the topic and a review of the literature is offered to discuss the parallel programming on multi-threaded CPU and GPU wireless cellular networks.

2.1 Evolution of Mobile Radio Access

A wireless cellular network is a one that has distributed base stations over a region each serving several mobile users connected to them. The communication between a mobile terminal and a base station is two-way, i.e. there is a dedicated uplink channel for mobile terminals to transmit as well as a downlink channel for base stations to transmit their data to the mobile users. In both uplink and downlink, the common channel is shared by multiple users. Multiple access schemes address the way the overall resources of the common channels are shared among the users within the network [23]. Chapter I

noted the cellular network's standards that have been implemented so far. The multiple-access techniques that have been used in those standards enabled the first voice call to be launched in 1981 which was based on FDMA. Nordic Mobile Telephone (NMT) system by Ericsson AB achieved a combination of several cells and released the world's first cellular network. The system remained analogue due to radio transmission technology even though the network had digital switching technology [24]

The cellular technology then experienced rapid developments and many digital technologies were implemented. In particular, the multiple access techniques over the years experienced the most dramatic changes. For example, in 2G, TDMA and FDMA were used in GSM. In the third generation cellular networks such as W-CDMA and UMTS, direct sequence CDMA (DS-CDMA) were used and the receiver employed the Rake receiver to counter multipath fading. Orthogonal multiple access (OMA) based on orthogonal frequency division multiple access (OFDMA) or its single carrier counterpart SC-FDMA was adopted in the 3.9G and 4G networks such as LTE [25] and LTE Advanced [26] [27]. These approaches rely on the idea that the common channel is partitioned and shared among the users either in time, frequency and code domain orthogonally, i.e., without overlapping each other hence the name 'orthogonal multiple access' (OMA). These technologies were successful to meet the demand for mobile services at the time.

In order to maintain the sustainability of wireless cellular networks during the next decade, novel solutions which will face the challenges are required [28]. Considering the 1000-fold leap in the size of mobile traffic, another advancements such as network capacity, quality of service and better user experience are needed, The cost-effective network capacity improvements are reached by smart design of radio access technology. Non-orthogonal multiple access (NOMA) scheme is proposed to accomplish those goals.

NOMA implements a novel approach for multiplexing users. Unlike in the previous generations of cellular networks, NOMA multiplexes users in power domain and let them operate in the same frequency and time. In NOMA, users are de-multiplexed at the

receiver side using advanced interference cancellation techniques such as a successive interference canceller (SIC) [29], [30]. From the perspective of information theory, NOMA with ideal SIC is an optimal multiple access technique in terms of achievable user data rates both in uplink [31] and downlink channels [32].

2.2 Basics Concepts of NOMA

This section introduces NOMA in the power domain and discusses it as a candidate of 5G. Interest around NOMA trend can be recognised in academic journals and conference proceedings approximately from 2012. The study continues both in academia and industry. As highlighted earlier in this paper, multiple access schemes share resources so that users in the cell simultaneously and actively connect to the base station. Massive usage of mobile devices and the popularity of the internet lead to a shortage of these natural resources such as frequency and time. NOMA in the power domain with an accurate interference cancellation at the receiver is proposed as a solution for both downlink (Fig. 2.1) and uplink channels (Fig. 2.2). From an information-theoretic perspective, NOMA with SIC is an optimal multiple access scheme from the view point of the achievable multiuser capacity region, in the downlink [32] [33] [34] and in the uplink [31]. In studies related to practical cellular networks [13], [30], again NOMA is more spectrally efficient when compared to orthogonal multiple-access techniques.

The benefits of NOMA does come with its challenges. Firstly, there is a need of an acute and reliable interference cancellation at the receiver. The lack of such will lead to error propagation and further deteriorates the performance of the NOMA receiver [35]. Secondly, since the user signals are distinguished in the power domain, the power allocation among the users is essential for improved capacity. Even with perfect interference cancellation, if the power is not allocated among the users properly, the expected performance outcome will not be reached. The efforts put on both the interference cancellation and power allocation are expected to pay back, however, these are usually sophisticated

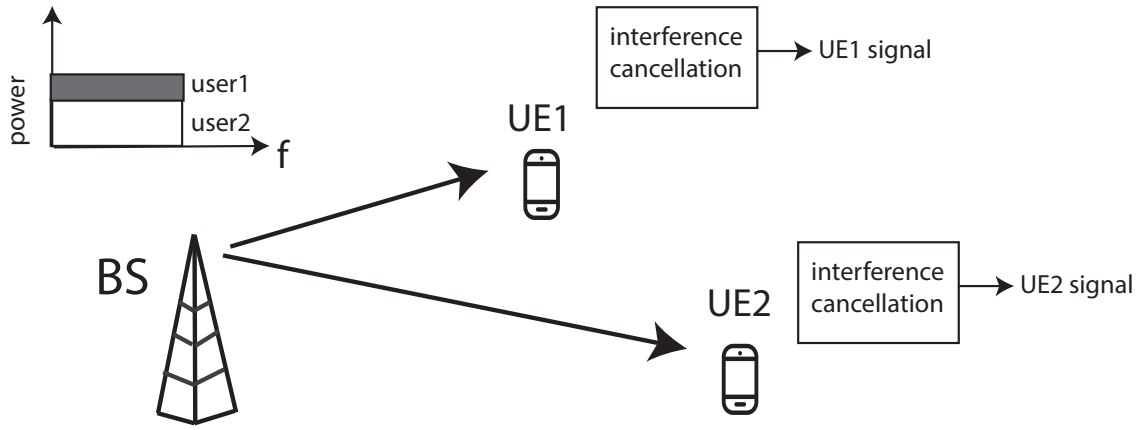

FIGURE 2.1: NOMA in power domain with one BS and two UEs in downlink channel.

and computationally heavy algorithms. Thirdly and often the challenge most overlooked, NOMA algorithms should run fast enough to meet the ultra-latency requirement of 5G networks.

2.3 Interference Cancellation

In downlink NOMA (see Fig. 2.1), the signals of each user are scaled and added on each other at the BS and then transmitted. The same superimposed signal is received by each UE in the cell. In uplink NOMA (see Fig. 2.2), on the other hand, each UE transmits their signal at the same time and at the same frequency. The signals are scaled in proportion to

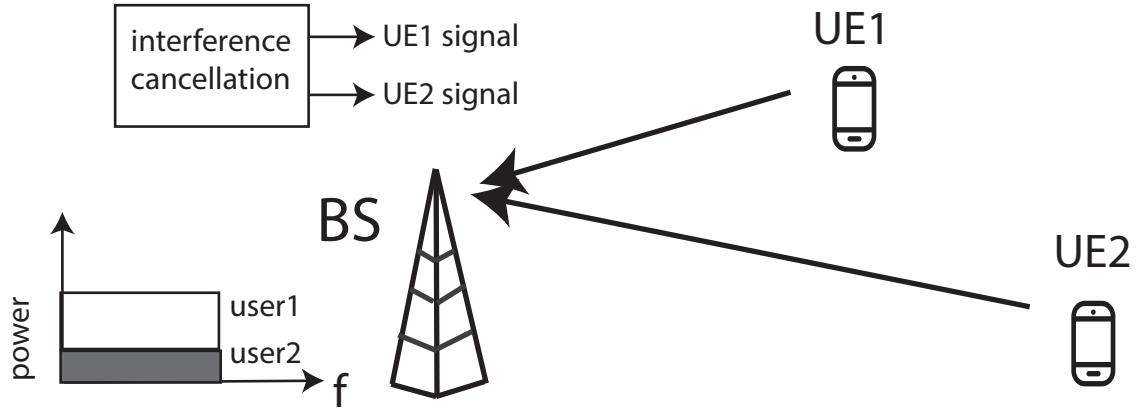

FIGURE 2.2: NOMA in power domain with one BS and two UEs in uplink channel.

their distances to the BS and added on each other when they reach the BS. In both uplink and downlink, to extract the desired signal from the superimposed signal, interference cancellation is applied. In this Section, the two most commonly used interference cancellation schemes in NOMA are introduced, successive interference cancellation (SIC) and parallel interference cancellation (PIC). Both the scholarly and industry literature on these techniques are discussed.

2.3.1 Successive Interference Cancellation (SIC)

Fig. 2.3 illustrates the working principle of a SIC receiver. The receiver decodes the received signal and the first decoded signal belongs to the one with strongest power. The remaining signals are seen as interference to this signal. The decoded data is then used to regenerate the signal as it was travelled through the channel. This can be done by modulating a carrier wave, multiplying with estimated channel coefficient and adjusting its phase. The regenerated signal is then subtracted from the received signal and the processes are iterated until the desired signal is obtained. The BS needs to iterate the process as many times as the number of users to decode each of them, whereas, UEs will iterate until they find their signal.

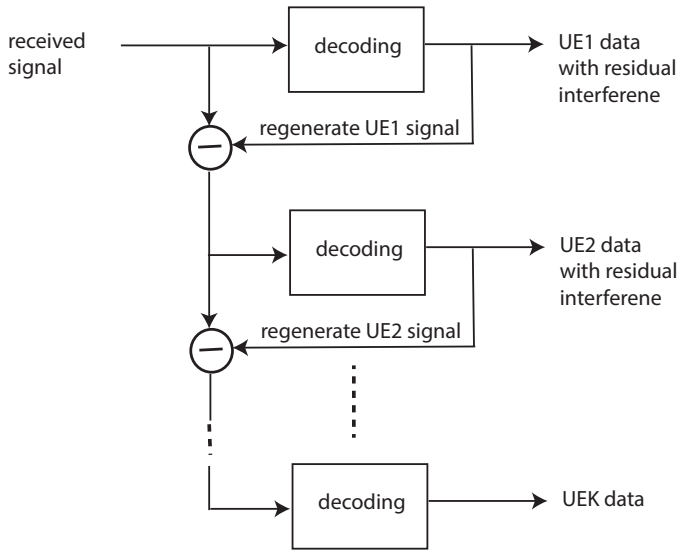

FIGURE 2.3: NOMA with SIC receivers for K UEs with channel gains: UE1 > UE2 > .. > UEK.

First versions of SIC were mentioned as early as 1994 [36] where SIC was discussed as a candidate scheme for direct sequence CDMA (DS-CDMA) networks with different and equal power distribution. The work implies that the more SIC iterations occur, the better BER (bit error rate) performance is achieved. Nevertheless, there was a tremendous decoding complexity due to the number of iterations. The work demonstrates this by analysing BER for 60 UEs without implementing complexity reduction techniques, such as user pairing, and finally compares the results of SIC with the conventional interference cancellation techniques.

Even though there are exceptions as in [23], where the iteration order starts from the weak signals, most of the recent works mentioned below employed decreasing order. For example, the strongest signal is decoded and cancelled first. The first decoded signal has the most amount of interference, while the last signal only decodes its signal and in case of perfect interference, the cancellation will experience only a noisy channel [37]. The SIC receiver has also some drawbacks that can be listed as

1. The scheme is dependent on the decoding order and substantial power difference is required.

2. The decoding time is proportional to the number of UEs in the network which is still expected to be compensated by the computation power of new mobile CPUs.

3. The channel information is needed at the receiver to prevent residual interference that leads to capacity erosion.

4. Interference that occurs from the neighbouring cell.

5. The errors made in decoding propagates to other iterations.

In the context of 5G, NOMA with SIC was first mentioned in a technical document by NTT Docomo [38]. In their works, engineers envisage the first weakness in the list above and kept their discussion with only two users in a cell. Soon after NTT Docomo's initiation, their works were extended to applications with multiple-input multiple-output

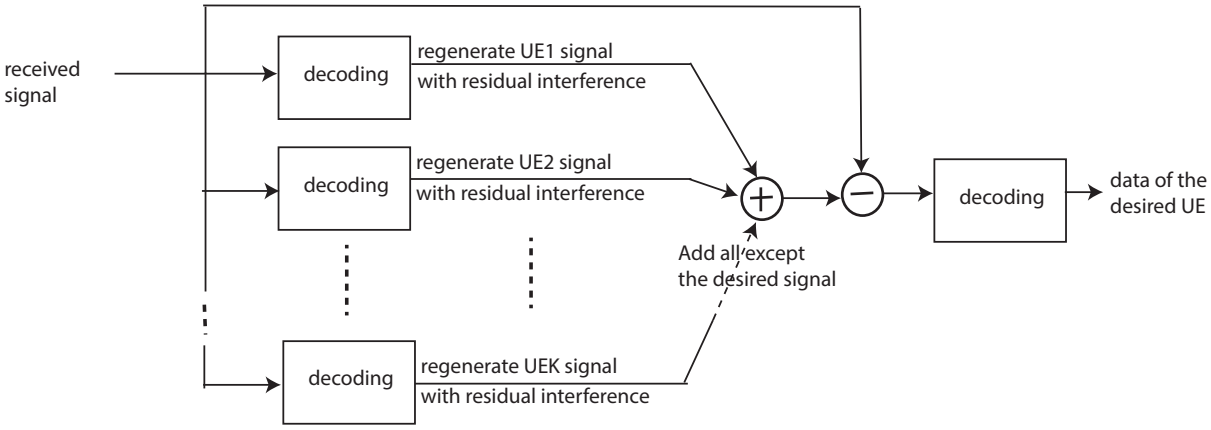

FIGURE 2.4: PIC receiver model for K UEs.

(MIMO). In [39] the data throughput of 'NOMA with SIC maintained by advanced power control' was compared with that of OMA and demonstrated significant performance improvement with NOMA. BER results for three UEs were more deteriorated compared to user pairing.

2.3.2 Parallel Interference Cancellation (PIC)

In this part, works related to PIC receivers with NOMA and CDMA schemes will be summarised. In the literature, there are fewer works for PIC compared to SIC, even though the scheme has considerable advantages over SIC. Furthermore, the performance of PIC may be enhanced via implementing it on recently available hardware upgrades such as parallel programming. PIC model for K UEs is illustrated in Fig. 2.4. In PIC receiver, all the signals except the desired one are demodulated in parallel, regenerated and then summed. The sum is then subtracted from the received signal along with the assumption that the difference is the desired signal. The final step is decoding the desired signal.

In comparison with SIC, in the literature, there are contradicting research examples that are biased to one of the schemes. For instance, BER results in [40] with up to 16 UEs showed a worse PIC performance than SIC where the comparison was made in both uplink and downlink channels. The study added OFDM computations to CDMA scheme and

proposed the receiver for multi-carrier CDMA. Results compared minimum mean square error (MMSE) per carrier, maximum ratio and equal gain combining receiver methods for single-user detection with hard and soft-input decoding methods.

Pioneers of PIC for 5G, Anwar et al. [41], revealed the drawbacks of SIC in downlink and proposed PIC as a more feasible alternative. The works discussed in [39] and [29] criticise SIC for its requirement for a substantial difference in the power level of each received signal, whereas in practice, this may not be maintained easily. Moreover, most works with NOMA with SIC note its dependency on the accurate decoding at the SIC iterations, otherwise it leads to error propagation [42].

The proposed PIC scheme solves the aforementioned problems [41]. The study demonstrated the probability of bit error with SNR and with the number of UEs, then computational complexity is presented. The study analyses the performance and admits the necessity to practical implementation of channel estimation. In their comparison of computational complexities for different number of UEs they find linear dependency for PIC scheme against $\mathcal{O}(logn)$ for SIC. Authors expect powerful smartphones in the next decade to cope with the tasks in parallel.

In this thesis, we evaluate the computation performance of both SIC and PIC receivers on different computing platforms in Section 3.3 and for multicarrier transmission NOMA-OFDM in Section 4.2. For SIC, the impact of decoding mismatch on the computation performance in terms of execution time is also discussed in Section 3.2.

2.4 Power Allocation in NOMA

In power domain NOMA, since the users are distinguished by their power levels, proper power allocation among the user signals is needed. As mentioned earlier in Section 2.3.1, for the SIC receiver to properly cancel the interference, the power levels of each contributing signal should be well differentiated. In downlink, the BS allocates its available power among the UE's modulated signals before adding them up. It allocates more power

the UEs that are located far from itself and less power to the ones that are closer to the BS [13, 29, 43]. This enables the SIC receivers to accurately cancel the interference. For instance, the farthest UE has the largest component in the received signal (the other signals seem as interference) and the in the first decoding iteration, the UE obtains its signal.

Whereas in the uplink, although the adjusting transmission powers of the geographically distributed UEs is possible with additional signalling, usually the difference in the received power levels of the UE signals is achieved naturally by their distances to the BS [44]. This time, assuming that all the UEs have the equal transmit power, the closest user would contribute more to the received signal and is decoded first at the SIC of the BS. Both in downlink and uplink, optimum power allocation remains as challenging problem due to its computational complexity and user mobility while satisfying many constraints such as latency and fairness in the network [37, 45–49].

In PIC, on the other hand, the receiver performs better than SIC when the received power level of each signal is equal [50]. This can be achieved by a power control mechanism as in CDMA networks [51] [50]. For PIC to work properly for a large number of UEs, a signature code is required to distinguish the user signals and the receiver becomes subject to code domain NOMA [52]. In the uplink, power control is implemented to orchestrate the UEs transmit powers so that the closer UE will have less transmit power and the further UE will have higher transmit power. In the downlink, the BS allocates more power to the signal of UE that is the furthest [53]. This power control strategy is routine work and had already been implemented in cdma One and 3G networks.

2.5 User Clustering in NOMA

Andrews and Meng had optimistic views [54] on the future radio networks of that time. However, their assumptions like *"only one decoder for all the users"* or *"Further as data-rate demands increase in the future, the number of users per cell will remain constant or*

possibly decrease." do not match to realities that multiple decoders may now serve the same cell. Moreover, the number of UEs per cell was much larger than they expected.

User clustering was a novel NOMA technique that deals with a large number of UEs by grouping them to diminish the complexity or to increase the scope of the network (see Fig. 2.5). User clustering was first mentioned for efficient optimal power allocation in [55] and several works followed on from the idea. In the related literature, user clustering is investigated for both uplink and downlink channels for several decoding schemes and power allocation algorithms.

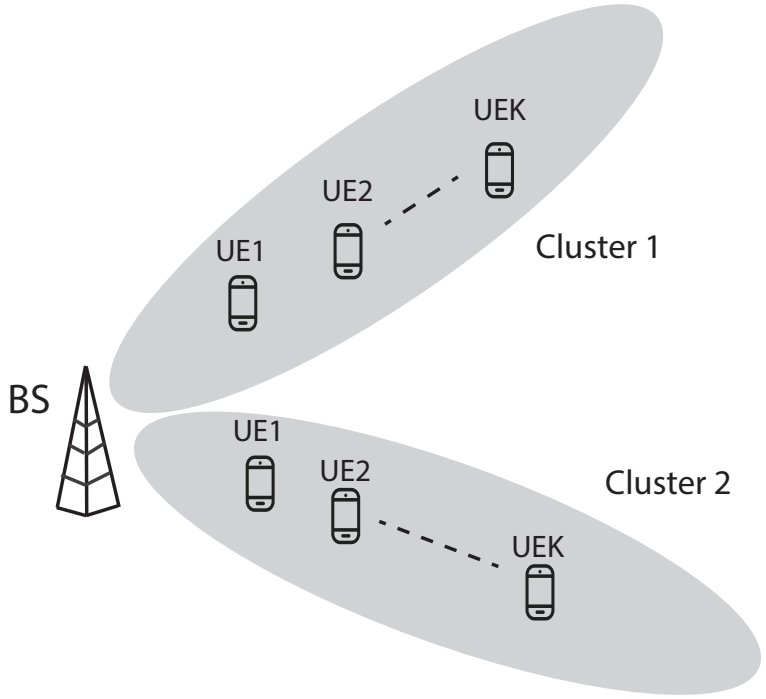

FIGURE 2.5: Cluster based NOMA system

Cluster formation, however, is a challenging task since different grouping strategies may give very different network performance. Most works in the literature consider a basic cluster with two users which is more like pairing rather than grouping the users. One major limitation of large cluster size is that it results in increased complexity and delay in SIC decoder. In [56], the two objective of cluster formation is listed as follows: 1) finding the optimum cluster size and 2) matching the users to clusters to maximise the sum capacity. The cluster formation is coupled with optimum power allocation and proportional fairness subjects as discussed earlier in Section 2.4.

In NOMA related literature, user clustering found its place in earlier works. For two UEs in downlink NOMA system, in [57] power allocation is studied using cognitive radio perspectives. [58] derives a suboptimal solution for max-min resource allocation and [59] studies the optimal resource allocation for maximising the sum capacity. Joint optimization of beamforming parameters and power allocation is studied in [60]. For the uplink in [61], user paring is studied and resource allocation is performed among these pairs. The aid of user clustering in full-duplex communication is another challenge for NOMA systems which is also investigated in [62]. Joint optimization of beamforming and MIMO parameters while detecting the users to be paired is proposed in [63]. User clustering is still an ongoing research field. The problems of optimal resource allocation and finding the best group are still be investigated. In this thesis, we approach the clustering problem from a computing perspective and propose parallel processing to challenge the computational difficulties in NOMA receivers in Section 4.1.

2.6 Parallel Programming

This section switches from wireless communications to the computer science field. It discusses articles parallel programming, multi-threading on CPU and then multi-threading on GPU. Furthermore, works which offer acceleration of wireless cellular networks with parallel programming concept are presented.

Future radio access is aiming for the simultaneous connection of multiple devices in one domain. The same concept is applied in parallel processing, where CPUs are multiplied physically and their tasks are assigned on multiple threads. Engineers keep trying to enable running more tasks concurrently and in parallel as possible. Physical resources such as hardware, special instructions and software are required to use the privileges of multi-threading. The hardware for running instructions is CPU which allows being instructed on a rich variety of programming languages. For example, java programming language allows direct instructions upon initiated virtual threads. Each thread executes a

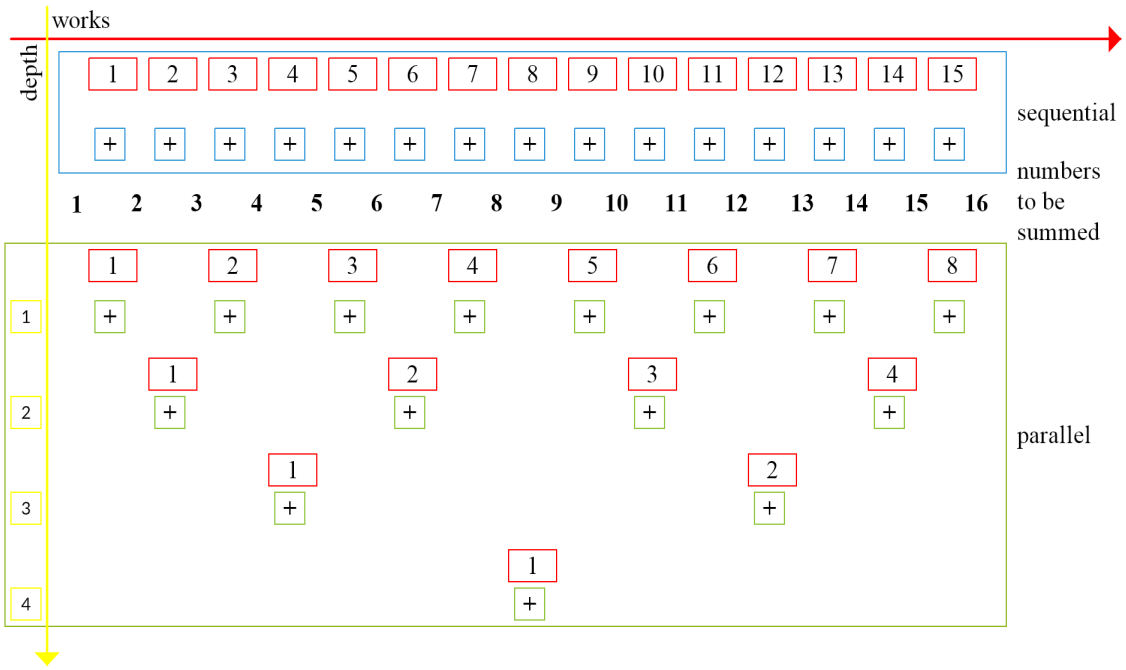

FIGURE 2.6: Sequential and parallel programming for summing numbers from 1 to 16.

particular assigned task and there is a possibility to initiate n number of threads. Multiple physical cores of each CPU allows it to initiate and stop a thread at a time.

The efficiency of parallel algorithms with work and depth terms are described in [16]. The term 'step' substitutes the word 'depth' in the latest studies which define the longest sequential chain of computations. The term 'work' stands for a number of all operations. Fig. 2.6 illustrates sequential and parallel algorithms for summing numbers from 1 to 16. In sequential summing, there are 15 works which run after another, whereas in parallel programming the works have another structure. Each depth has half as many works as the previous level until there was only one summing work left. The idea is that those summing operations are independent and maybe run on different processors, cores or threads simultaneously.

There are early works that discussed computation requirements for hardware and also considered CPUs in the BS [64], [65]. The study revealed the scarce of hardware possibilities for updated algorithms of those times. The authors compared CPUs of the BSs, which were used for GSM e.g. in the 1990s. Intel® Pentium 166, SUN Ultra 170 and

DEC 3000/80 processing units were compared. Authors analysed each phase of the receiver such as channel coding, interleaving, modulation and detecting errors separately. The simulations were done with C++ programming language which was modern at the time. Both works expect CPU accelerations for modules. Interestingly, computations in [64] had foreseen clustered workstations or multiprocessor server appearances and became widely available in 4-5 years. Those workstations had to run computations similar to theirs. The works also introduced a method to benchmark algorithms and were done on MATLAB software.

About a decade later a study proposed multi-core digital signal processor architecture for cellular networks [66]. A discussion included debugging of parallel codes and Field Programmable Gate Arrays (FPGAs) and even obsolete Application Specific Integrated Circuits (ASICs). The work only proposes conceptual architecture without a practical discussion. Then multi-core physical layer for the super base station was proposed for time-domain long term evolution (TD-LTE) in [67]. Proposed digital signal processor for data control, uplink and downlink data processing were located on multiple cores.

Another work proposed speed enhancement via parallel computations in multiple access MIMO channels [68]. CPU cores or threads were not involved in the offered alternative optimization algorithm. Results presented the average worst-case MSE of the proposed transceiver design. Design of the proposed robust transceiver outperformed non-robust transceiver one. There are several works which modify algorithms and implement parallelisation in information theory [69–72]. These show the benefit of parallel concept became recognisable in different layers of wireless communications and availed in the design/implementation of both hardware and software.

The number of cores in conventional CPUs nowadays reaches up to 18, whereas the graphics processing unit has built-in thousands of cores. Those GPU cores are also available for general-purpose programming along with graphics visualisations. Thus, GPU solves problems with a large amount of data or many similar small-sized tasks in a relatively short time. Open CL by AMD and CUDA by NVIDIA® are software platforms for

GPU acceleration. CUDA is being widely used in machine learning, computational fluids dynamics, industries which use weather/climate data and generally applied in research of different directions [73]. Telecommunications are also becoming popular and LTE BS is based on GPU as proposed in [74]. After DSP and FPGA based BSs, the new computation hardware is offered and the results are compared with the benchmark. The study involved two GPUs and concluded that in a real-time scenario more GPUs are required. These are the benefits of GPUs highlighted from the articles:

1. GPUs have thousands of threads and offer general-purpose programming
2. Data-level parallelism.
3. Threads parallelism.
4. Consume less power and therefore cost-efficient.

Concluding the benefits listed above, the article finds GPU as the best hardware alternative for an LTE BS. Authors suggest allocating different threads per connected UE. Antennas, symbols and subcarriers may also be served by GPU threads in parallel. CPU is needed to operate GPUs. The researchers used two GPUs and reached the data rate of 75 Mbits/sec in a setup of the article. Another recent study applied the CUDA platform to accelerate complex computations in massive MIMO systems [75]. The GPUs are not widely implemented in accelerating problems of wireless communication. Neither the algorithms of computation are discussed. NVIDIA® is researching 5G and has a special direction of telecommunications in famous annual GPU technology conference [76], [77].It may be concluded that there is a credible premise for GPUs in 5G which comes naturally rather than a crammed mixture of technologies. This work aims to show some practical unbiased appliances of parallel programming on GPU for NOMA. Indeed, NOMA with related procedures is expected to be applied in wireless communications cellular networks of the future generation.

2.7 Artificial Intelligence in Future Networks

The aim of AI assistance in wireless communications is that upcoming mobile networks can learn several system parameters from users' behaviour to the changes in physical layer characteristics to autonomously optimise itself for improved performance. This smart approach of self-configuration requires the implementation of sophisticated AI tools. However, ML and more recently proposed DL tools have already found their places in the context of AI-assisted wireless communication, though it is more conceptual than practical today. In Fig. 2.7, the illustration shows a smart base station that learns through observations and then executes actions based on its evaluations, i.e. optimisation of selected system parameters.

ML techniques have been considered for wireless systems to estimate massive MIMO channels, user location learning, spectrum sensing, device-to-device user clustering, resource allocation, spectrum sharing, energy harvesting and HetNet selection (see [78] and the references therein). DL was also incorporated into wireless communications in the context of channel coding for MIMO in [79–81]. In [82], DL was used to develop a codebook adaptively to increase the error performance of the SCMA systems. The authors also demonstrated the superiority of DL-aided methods in terms of computational time. AI has also been applied to multicarrier systems for channel estimation and signal detection in [83] and its ability to detect signals directly as opposed to conventional receivers which estimate channel first and then detect the signal. Moreover, DL aided data traffic

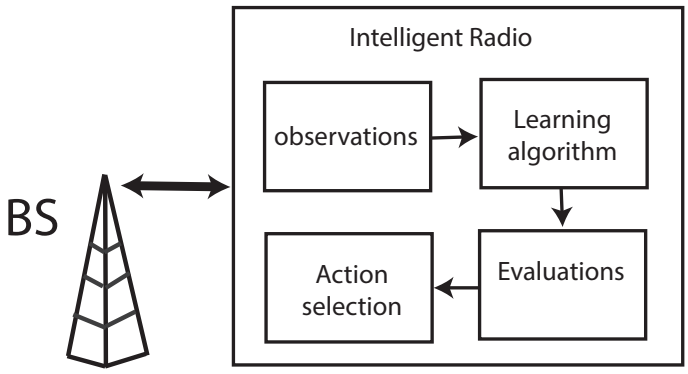

FIGURE 2.7: Illustration of learning platform of AI assisted radio.

management was considered in [84–87].

The recent popularity of AI and useful assistance from ML and DL results in a plethora of works. It can be seen that advancements in algorithms barely challenge with the non-linearity rules and policy obtained with the AI. In Chapter 5, the power allocation problem in the downlink NOMA predicts the power allocation coefficients to maximise the sum-rate by applying AI algorithms is explored. Although many other recent papers have also attempted sum rate and reliability optimisation, these methods require high computational complexity due to the nonlinear optimisation. This paper further compares the computation times of AI-aided techniques with optimal numerical search methods and demonstrate superiority.

Chapter 3

NOMA with SIC and PIC

This chapter deals with optimum power allocation mechanism for NOMA networks (Section 3.1), improved security thread detection for NOMA with SIC (Section 3.2) and a method which decreases the running time of PIC and SIC receivers during decoding the signal (Section 3.3). These solutions are meant to fill the gaps identified earlier in the previous chapters. Specifically, the literature referring to power allocation in NOMA mostly consider applications with user pairing. We first demonstrate optimal power allocation with exhaustive search algorithm that is meant to determine the feasibility of the idea and to build a guideline for our research. Moreover, the algorithm keeps a balanced throughput ratio among UEs via predefined fairness index. Since there is scarce attention paid to the NOMA security in the literature, a thread detection that takes into account the execution time of the SIC receivers is proposed. Finally, we present our numerical results related to the execution time for SIC and PIC on multi-threaded central and graphical processors.

3.1 NOMA Optimum Power Allocation

3.1.1 Introduction and Related Works

Power allocation is an open research field of significant interest for NOMA with SIC. In the literature, there are many power allocation techniques to increase the overall spectral efficiency. Ideal power allocation is exceptionally challenging since it also determines the

interference level in the network. The problem gets more complicated for multicarrier transmission. Earlier attempts on power allocation for multi-carrier NOMA systems, for example [29], [88], [30], are far from being optimal. The works in [49], [89] consider only two UEs and reduce the number of subcarriers. Then they can propose optimal resource allocation for a hybrid orthogonal and non-orthogonal access scheme. Other works [90] incorporated subcarrier allocation and obtain close to optimum performance with convex programming but again for two users only.

In optimal power allocation, the objective is usually to maximise the sum rate of the network. The optimisation becomes a non-convex problem due to the interference term in the objective function. This research [91] proposes an analytical suboptimal solution for power allocation that maximises the sum rate by employing a water-filling algorithm. In [92] authors defined a closed-form suboptimal solution for joint subcarrier and power allocation for multicarrier NOMA system combining Lagrangian duality and dynamic programming. Most of the works in the literature consider maximising the sum throughput, however, fairness is another important objective to be considered in resource allocation. There is a trade-off between the fairness and the sum data rate in cellular networks. Optimal resource allocation solutions incorporated different fairness measures to their problem definitions in early works such as [44, 45]. In [93], authors derive a closed-form optimal solution for multicarrier NOMA with fairness constraint for two users available in the network. The recent study in [93] proposes proportional fairness performance with and without user pairing scenarios for power allocation in NOMA uplink. The work involves various decoding orders for both cases. However, the algorithm complexity of both scenarios is $\mathcal{O}(N^2)$, which is time-consuming comparing to works previously described. In [94], power allocation is proposed for two and three multiplexed UEs with Karush-Kuhn-Tucker conditions and the performance is compared with exhaustive search. The results are obtained considering the 3GPP scenario. On the one hand, the algorithm has $\mathcal{O}(N)$ complexity and aims to maximise sum capacity rather than balancing the ratio or max-min rate of each UE.

In addition to the fairness index, max-min methods are also employed to demonstrate fair scheduling. For example, [95] employs power allocation for NOMA by maximising the minimum data throughout in the network using channel state information. Their problem is non-convex, however, the proposed solution is low-complexity and provides a close-to-optimal resolution. Research that considered max-min fairness in [96] performed power allocation for outage balancing problem for downlink NOMA that maximises the minimum outage probability.

Due to the computational complexity and challenges in fairness, recent works concentrated more on user pairing within the cell. The articles [97–100] discuss the NOMA scheme with SIC receiver for user pairing. The works include interference rejection combining algorithm with MIMO system; optimum against random user pairing scenarios; different number of antennas at the BS and the UEs side without inter-cell interference as well as compare user pairing with a tree search based transmission algorithm.

3.1.2 System Model

In our system model, a downlink transmission was considered in the wireless cellular network with NOMA scheme in the power domain. The network has a BS and K UEs each having SIC receivers. The distance of each UE to the BS is different. UEs are ordered starting from the UE_1 as the closest one to the BS and ending with UE_K, which is the furthest from the BS (Fig. 3.1). Since the NOMA scheme is in power domain, the signal for UE_1 is assigned the least amount of power and thus has the weakest channel conditions. Whereas the channel conditions of the farthest UE_K are compensated with the largest amount of power. Therefore, UE_K has the strongest signal. As for the receiver side, all UEs receive the same transmitted signal with messages for all UEs and then each UE runs SIC for decoding. Received signals are decoded starting from the strongest one, then the decoded message is subtracted from the received signal. These decoding procedures and subtractions are iterated until the signal which needs to be decoded reaches its order at the SIC receiver. The nearest UE cancels out the decoded signals of all further UEs and

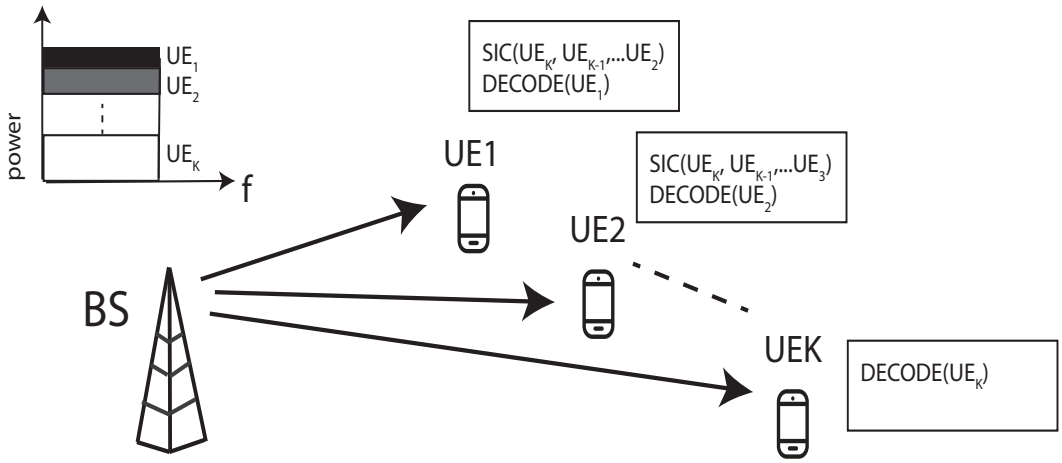

FIGURE 3.1: Downlink NOMA with SIC for K UEs

the farthest just decodes its signal due to the signal strength. The remaining part of the signal is considered as interference.

The BS modulates the message signals of each UE with a single carrier modulation scheme such as quadrature phase-shift keying (QPSK), quadrature amplitude modulation (QAM) etc. Then superimposed signal $x(t)$ (Eq. 3.1) is obtained by summing up of all the modulated individual waveforms. Finally, the BS transmits the signal (see Fig. 3.2 (a)), which can be written as

$$x(t) = \sum_{k=1}^{K} \sqrt{\alpha_k P_T} x_k(t). \tag{3.1}$$

As can be seen in (3.1), it is the sum of $x_k(t)$ messages of every UE scaled with the power allocation coefficient α for UE_k and the total power P_T is the available power at the BS. It turns out that the power of UE_k signal is $P_k = \alpha_k P_T$ with coefficient values assigned according to the distance as described earlier.

The signal received by each UE is written as

$$y_k(t) = x(t) g_k + w_k(t) \tag{3.2}$$

g_k is the channel attenuation between the BS and each particular UE_k, $w_k(t)$ is the additive white Gaussian noise at the UE_k with mean zero and power density N_0 (W/Hz) and $x(t)$

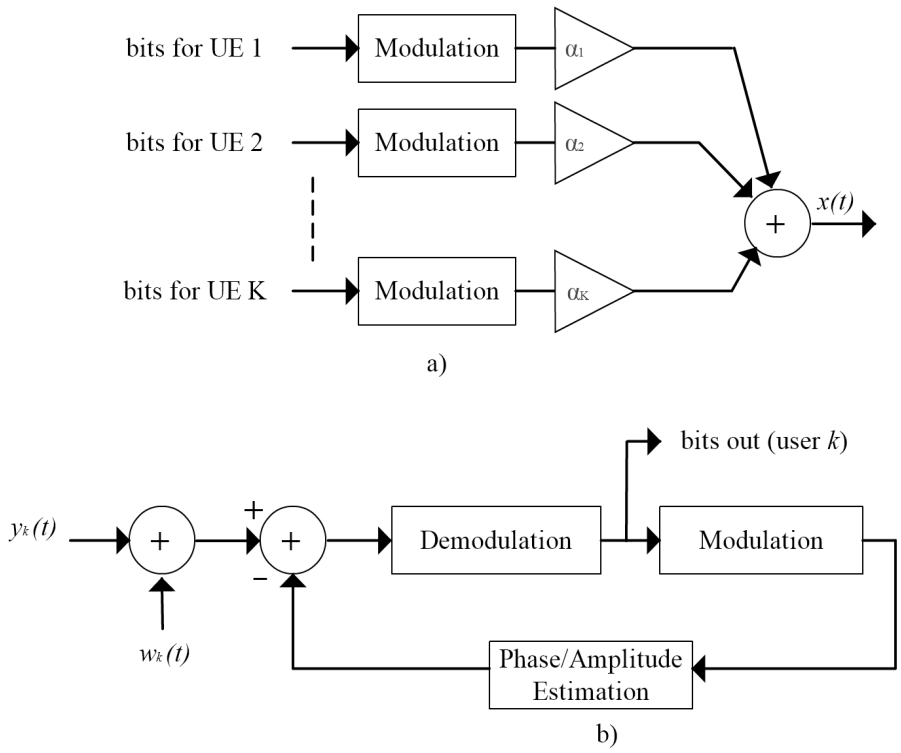

FIGURE 3.2: Block diagrams of downlink NOMA with SIC (a) transmitter side (b) receiver side.

is the superimposed signal.

The data throughput for NOMA can be represented as follows [23]:

$$R_k = W \log_2 \left(1 + \frac{P_k g_k^2}{N + \sum_{i=1}^{k-1} P_i g_k^2} \right) \tag{3.3}$$

where W is the transmission bandwidth and N is the total noise power $N = N_0 W$. For OMA the total bandwidth of the channel and power of the transmission are shared equally for all UEs, giving

$$R_k = W_k \log_2 \left(1 + \frac{P_k g_k^2}{N_k} \right) \tag{3.4}$$

where $W_k = W/K$ and $N_k = N_0 W_k$. The sum capacity is an aggregated result of all UEs is expressed as

$$R_T = \sum_{k=1}^{K} R_k. \tag{3.5}$$

The fairness index of our system model is defined by [101] as

$$\mathrm{F} = \frac{(\sum R_k)^2}{K \sum R_k^2} \tag{3.6}$$

which measures the fairness of capacity of each UEs. The value of F varies between 0 and 1. As higher the index as more adjacent becomes capacity values of neighbouring UEs.

Algorithm 1: Optimum power allocation (OPA)

```
initialization
initialize powerMatrix include all possible PAs
set fairnessConstraint to F'
for i in powerMatrix do
   calculate capacity
   calculate fairnessIndex
   if fairnessIndex ≤ fairnessConstraint then
      set capacity(i) to zero
   end if
end for
set maximumCapacity to zero
for i in capacity do
   calculate capacity(i)
   if capacity(i) ≥ maximumCapacity then
      set maximumCapacity to capacity
   end if
end for
```

The optimum power allocation aims at maximising the total capacity of connected UEs within the fairness index constraint.

$$\begin{aligned} \underset{\alpha_k}{\text{maximize}} \quad & \sum_{k=1}^{K} W \log_2 \left(1 + \frac{P_k g_k^2}{N + \sum_{i=1}^{k-1} P_i g_k^2} \right) \\ \text{subject to:} \quad & \sum_{k=1}^{K} P_k \leq P_T \\ & P_k \geq 0, \ \forall k \\ & \mathrm{F} = F' \end{aligned}$$

where F' is the targeted fairness index. Capacity is measured along with the fairness index. Each UE_k is assigned a power allocation coefficient α_k, which is obtained from the

exhaustive search described in Algorithm 1.

3.1.3 Numerical Results and Discussion

This section presents the numerical results which compare achieved data rates from the OMA and NOMA with different power allocation coefficients. The results were obtained with the following parameters (Table 3.1).

TABLE 3.1: Simulation Parameters

Parameter Name	**Parameter Value**
Number of UEs (K)	5
Bandwidth (W)	50 MHz
Total Power (P_T)	1 Watt
Distance between UEs	50 Meters
Carrier Frequency	1 GHz
Noise Density	10^{-17} W/Hz
Propagation Model	Okumura-Hata
Channel Gains for 5 UEs	[-33.21 -36.23 -37.99 -39.24 -40.20]
Fairness Index	0.9

Data throughput for OMA with equally divided power allocation and bandwidth obtained by (3.6) with a fairness index equal to 0.9 is given in Fig. 3.3. The UE_1 achieved the highest data rate of $9x10^7$ bps, the second one has close to $7x10^7$ bps, the data rate values decrease until it reached slightly more than $4x10^7$ bps for the UE_5. The sum capacity for all UEs in the OMA scheme equalled $3.05x10^8$ bps.

Compared to the results obtained with OMA, the NOMA evaluation of the data throughput with optimally allocated power using (3.3) was considered. Numerical results prove that NOMA outperformed OMA in all the selected three fairness constraints that were 0.5, 0.7 and 0.9. Data throughput results of NOMA with the lowest fairness index equal to 0.5 are imbalanced (see Fig. 3.4). The overall sum capacity roughly reached $4.37x10^8$ bps which was substantially higher than that of OMA. The first UE achieved $2.5x10^8$ bps and the second UE barely reached $1x10^8$ bps. Obtained optimum power allocation coefficients for all UEs become $[\alpha_1 \ldots \alpha_5]$ = [0.07 0.2 0.23 0.24 0.26].

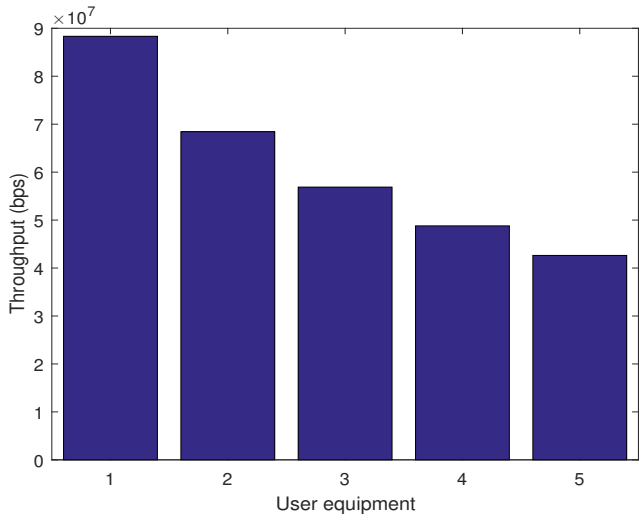

FIGURE 3.3: OMA maximum capacity data throughput rate performance for five UEs with 0.9 fairness index.

Fig. 3.5 shows the capacity for NOMA with a fairness index of 0.7. Farthest two UEs obtained less data throughput comparing to data throughput of OMA. However, the first three UE achieved more capacity. The highest data throughput 16 x 10^7 bps belonged to the first UE, and 12 x 10^7 bps had a UE 50 meters farther. 5 x 10^7, 3.8 x 10^7 and 3 x 10^7 bps were achieved by the last three UEs. Optimum power allocation coefficients for this capacity were $[\alpha_1 \ldots \alpha_5]$ = [0.02 0.14 0.23 0.30 0.31]. The sum capacity became 4.17 x 10^8 bps, which was still higher than that of the OMA and slightly less than the sum capacity of NOMA with a fairness index equal to 0.5.

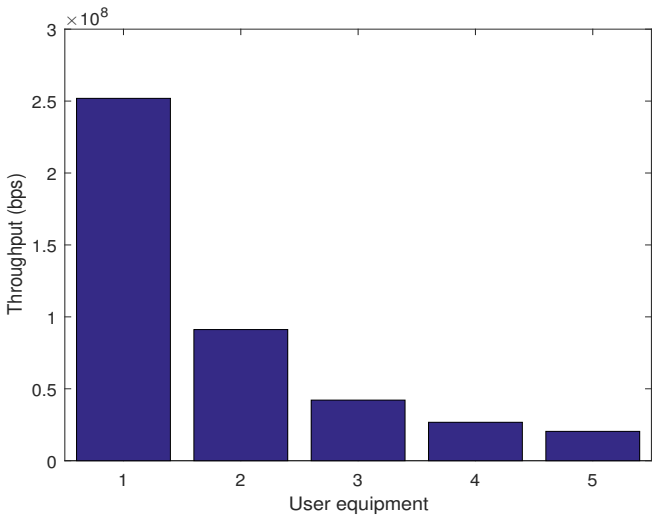

FIGURE 3.4: NOMA data throughput for 5 UEs with 0.5 fairness index.

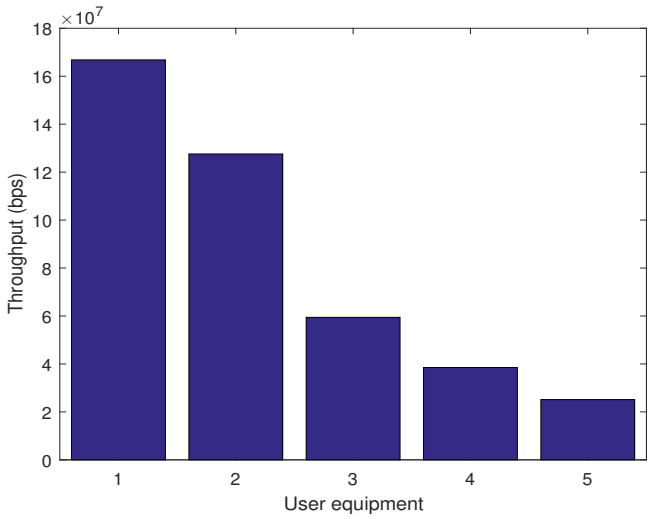

FIGURE 3.5: NOMA data throughput for 5 UEs with 0.7 fairness index.

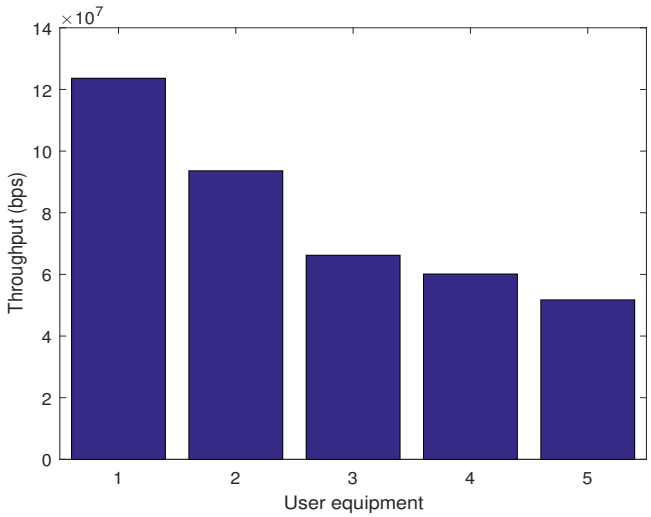

FIGURE 3.6: NOMA data throughput for 5 UEs with 0.9 fairness index.

The toughest fairness constraint of 0.9 significantly decreased data throughput of NOMA for 5 UEs (see Fig. 3.6). The sum capacity for all UEs became 3.95 x 10^8 bps, comparing to 3.05 x 10^8 bps for OMA. The furthest UE reached 5 x 10^7, whereas in the OMA worst results are 4.4 x 10^7 bps. The highest data throughput was achieved by the UEs closest to the BS in both schemes. In NOMA the first UE had only 0.01 allocated power coefficient. Data rates were 12.1 x 10^7 bps in NOMA and 4.4 x 10^7 bps in OMA for the first UE. In NOMA with OPA coefficients, all UEs became as $[\alpha_1 \ldots \alpha_5]$ = [0.01

0.05 0.12 0.28 0.54] when the fairness index was set as 0.9. Such data throughput values were close to real-life scenario provided by the 4*G* network [102].

In summary, the fairness index requirement is inversely proportional to the sum capacity. The higher the fairness index requirement was less the difference between the obtained data throughput of UEs. In all cases, data throughput decreases starting from the UE_1 until the UE_5. Last three UEs have a higher capacity as the fairness index is increased, whereas the capacity of the first UE decreases. Data throughput of the second UE gradually decreases starting from the 0.7 fairness index. Capacity ranges in between the 0.5 x 10^8 and 1.2 x 10^8 bps for the highest fairness constraint 0.9.

3.2 Security Threat Detection for SIC

3.2.1 Introduction and Related Works

Wireless cellular communication researchers' aim mainly to improve spectral efficiency whilst maintaining reliability. The industry expects the mastery of combining several existing technologies such as IoT, Big Data, AI etc. [103, 104]. The Next Generation Mobile Networks (NGMN) Alliance advises putting the priority on fundamental network quality and the security of future mobile networks [105]. There is a tendency of the research to be directed to improving the capacity, energy efficiency, latency and security. Trust, identity and privacy are vital concepts which form the security. There is a need for strong immunity to protect cells and keep the privacy of UEs [103].

The NOMA with SIC is one of the most discussed concepts for 5G along with cognitive radio, mm Wave, VLC and massive MIMO [30]. The downlink channel of NOMA with SIC assumes that multiple UEs in a cell receive the same encoded signal transmitted from the BS. Then each UE decodes the message following the order set by its power level and acknowledged by the BS [106]. Such iterative manner of decoding allows to access the message of others, i.e., the ones that does not belong to the UE. The study in [107]

offers to secure the privacy by MAC and IMEI (International Mobile Equipment Identity) hardware unique identifications to encrypt each UEs message in NOMA networks.

The study in [108] mentions some possible security issues caused by the natural weakness of wireless cellular network channels. The research work in [109] relates to the physical layer security in NOMA schemes. In [110], the authors propose a new architecture with external eavesdroppers for NOMA. Variations of orders to choose antennas in the SISO and MISO NOMA networks enhances the performance of security in [111]. The study in [112] raised the sum rate for the SISO NOMA cell with external eavesdropping. The work considers the quality of service constraint. Furthermore, [113] investigates the security in MISO NOMA via beamforming. The scheme increases the security level of private UE.

There are very few studies which demonstrate the security flaws in future wireless networks and particularly in NOMA with SIC. Thus, this section of the thesis is devoted to detecting a security threat of NOMA with SIC.

3.2.2 System Model

Besides optimum power allocation, this section addresses a security issue at the SIC receiver. Security of cellular networks is cared for each layer starting from the physical one. The NOMA scheme with the SIC receiver considers decoding the needed message from the superimposed signal. According to the SIC algorithm, UE_k executes particular number of iterations to decode the needed signal. Then it is obvious that time to decode the signal depends on the number of iterations. The number of iterations are defined from the order, which is established from the channel gain. From this we can infer, that the channel gain affects the time to decode the signal. It may be recognised that UE_a spends an 'unusually' long or short time to decode the signal. This is the hypothesis that UE_a is not following the correct decoding order and tries to decode others, such as UE_b. Then time to decode UE_a and UE_b are similar. In other words, if UE_a pretends to be ordered as UE_b, the receiver will decode the signal if the message addressed to UE_b. It may be

assumed that the cause of such 'anomaly' is an attack from a third part [114–117]. In the following section, the change in decoding time of an UE is discussed in the scenario where it attempts to hack other UE's messages.

In downlink NOMA with SIC, all UEs receive the same superimposed signal which includes messages from all connected devices. Then the receiver cancels out those that of other ones according to the decoding order. Both the transmitter and the receiver sides know the allocated power, channel gains and the number of messages per sent signal in advance. The decoding order determines the time spent on interference cancellation and then decoding the desired message. Thus, order mismatch might be a result of a malfunction, i.e., an attempt to decode the other UE's message. This may be recognised by an abnormal amount of time spent. For example, in normal operation, the UE_k needs to cancel signals of UE_K, UE_{K-1}, ... , UE_{k+1} and therefore, the rest UE_{k-1}, ... , UE_1 are treated as interference, so the malfunction leads to interference cancellation for UE_l where $\ell \neq k$.

3.2.3 Numerical Results and Discussion

The research's numerical results show a possible threat detection upon the SIC receiver. There are 13 UEs in the network and the power allocation coefficients are taken with the algorithm in Section 3.1 [118].The number of UEs in this simulation is higher than those in the earlier discussion. In the analysis, each UE has information on power allocation coefficients and the channel gains. The simulation parameters are given in Table 3.2. In the analysis, each UE has information on power allocation coefficients and the channel gains.

Incorrect decoding order for a particular UE was simulated by changing its decoding order for adjacent UEs. In this work, a selected UE_k was allowed to decode the signals of UE_{k+1} or UE_{k+2}. Such an attempt to decode other higher-order UEs takes less time than decoding the signal of its own. It is also possible to simulate the case that UE_k decodes UE_{k-1} which leads to the expenditure of more processing time.

TABLE 3.2: Simulation Parameters

Parameter Name	**Parameter Value**
Number of UEs (K)	13
Trials	100 000 Times
Bandwidth (W)	50 MHz
Software Parameters	
Software	MATLAB (tic-toc)
Hardware Parameters	
CPU	Intel® Core i7
Physical Cores	4*2.3 GHz
Memory (RAM)	4 GB 1600 MHz DDR3
Modulation	QPSK and 64-QAM

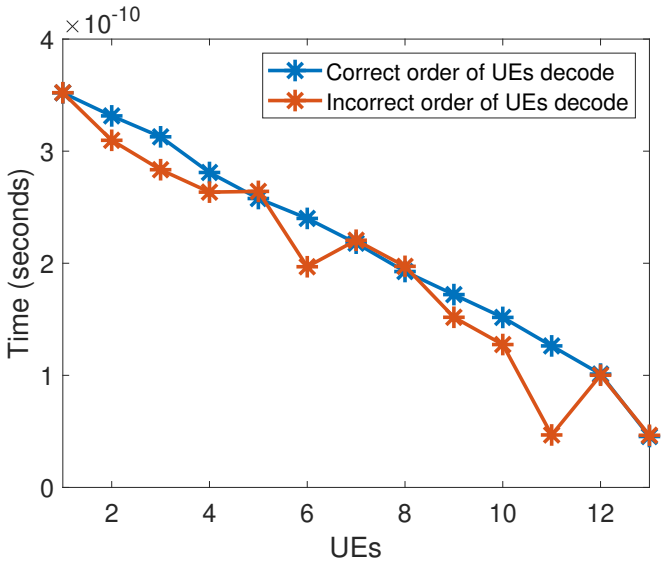

FIGURE 3.7: SIC computation time for 13 UEs.

Fig. 3.7 shows the SIC computation time for 13 UEs. The figure depicts the spent time for SIC execution with correct and incorrect decoding orders. The correct order is the case when each UE decodes their own signals without any mismatch. However, in the incorrect order, the UE_6 and UE_{11} have mismatched decoding and also some fluctuations are observed. These fluctuations are the result of their attempts to decode the signals of other neighbouring UE. Therefore the time is abnormally less than what it supposed to be in the correct decoding order. The decoding time has a decreasing flow starting from the $3.5x10^{-10}$ for the UE_1 and reaches about $0.5x10^{-10}$ for the UE_{13}.

In the analysis, the first UE is the closest to the BS and needs to decode all the other

signals iteratively to obtain its message. Whereas the signal of UE_{13} is allocated more power, as it is the farthest one and it only decode its message. The red curve in Fig. 3.7 stands for the incorrect decoding order for the UE_6 and UE_{11}, e.g., UE_6 attempts to decodes either UE_7 or UE_8. As it can be observed, even one less iteration cycle can be detectable.

Another point is that the decoding order mismatch may not be necessarily due to an intentional attempt but can be caused by hardware-software impairments. In any way, decoding order mismatch may lead to re-request of the signal by the UE from the BS, e.g., via automatic repeat request protocol and repeating the SIC procedure. The computation time allows us to estimate normal and abnormal time to recognise the mismatch and that the communication has not performed well in the first case. A possible solution is to keep an eye on the average decoding times of the UEs and track the number of re-requests or negative acknowledgements, either to detect the threat or receiver malfunctions.

3.3 NOMA with SIC and PIC

3.3.1 Introduction and Related Works

NOMA scheme is widely discussed as a candidate for multiple access scheme for 5G in both uplink and downlink channels. The scheme superimposes the transmitted signals in the same frequency and time, but they are distinguished by the difference of their power levels. As we discussed earlier, SIC and PIC receivers are the most commonly used two interference cancellation techniques in NOMA. The former one removes the multiuser interference with the cost of error-propagation and time delay, whereas the latter one neglects the power difference [14, 41, 43].

The SIC suppresses the interference by iterative decoding the signals. However, this iteration causes another limitations upon the performance of the receiver like error propagation and increased latency. There are studies that offer modified and enhanced variations of SIC to avoid error propagation. For instance, [119] proposes a sub-optimal multi

UE identification, then [120] investigated outage equalisation for the UEs decoded after a particular step via advanced SIC planning. Dynamic SIC order is figured out based on the channel information in [121]. The research in [122] claims that advanced iterative cancellation may work out for NOMA.

PIC alternatively solves the problem of error propagation. The receiver approximates the signals of all other UEs in the network before subtraction (or cancellation step) to obtain its signal [123–125]. For a large number of UEs, however, it requires signature codes and power control mechanism as in CDMA networks.

To conclude, there are lots of studies that propose SIC receiver for NOMA and fewer works that consider PIC as a promising receiver with more emphasis on avoiding error propagation and improving the reliability. However, to the best of our knowledge, there is little published work that covers the computation time of both receivers. Moreover, there are no studies that demonstrate SIC and PIC receiver implementation on GPU and also multi-threaded CPU.

3.3.2 System Model

In this section, the performances of SIC and PIC receivers are compared for uplink NOMA. According to our scenario, K UEs simultaneously send signals to one BS in the cellular network. The received signal at the BS is represented

In this section, the performances of SIC and PIC receivers are compared for uplink NOMA. According to our scenario, K UEs simultaneously send signals to one BS in the cellular network. The received signal at the BS is represented

$$y(t) = \sum_{k=1}^{K} \sqrt{P_T} x_k(t) g_k + n(t) \tag{3.7}$$

where $x_k(t)$ is the modulated waveform transmitted from the UE_k, g_k is the channel attenuation coefficient between the BS and the UE_k, P_T is the total power at the UE and $n(t)$

is the additive white Gaussian noise (AWGN) with zero mean at the BS and N_0 (W/Hz) density.

BS implements either SIC or PIC to the received signal. In the case of SIC, its working principle was described in Section 3.1 and illustrated in Fig.3.1 and Fig.3.2 (b), as for the PIC, it is now illustrated in Fig. 3.8. The SIC process in the uplink channel is similar to the SIC in downlink channel, with the difference that the receiver decodes each signal and subtracts them from the received signal iteratively. The UE with the highest power, i.e., the closest to the BS, is demodulated first and the other UE's signals remain as interference. In the second iteration, the UE with the second-highest power is decoded. These steps are continued until the last UE. Bandpass variant of the demodulated signal with correct phase and amplitude is imitated in every iteration to be subtracted from the received signal.

PIC receivers can decode all the UEs simultaneously regardless of their power levels. After decoding, estimated messages are summed apart from the message of the UE_k and subtracted from the received signal (Eq. 3.8). Finally, the resultant signal is decoded and the desired kth message is obtained.

$$\begin{aligned} &x_k(t) = y(t) - \textstyle\sum_{\substack{k=1 \\ k\neq i}}^{K} \hat{s}_i, \\ &i = 1,2,...,k-1,k+1,...,K. \end{aligned} \tag{3.8}$$

3.3.3 Numerical Results and Discussion

This section presents the comparison of the computation times of SIC and PIC receivers for various number of UEs. Simulation parameters are given in a Table 3.3.

Figs. 3.9, 3.10 and 3.11 show the computation times of SIC and PIC on different platforms. Results of both modulation schemes have approximately the same execution times in all three environment setups. However, the behaviour of the curves also depends on the number of UEs and the receiver type. Results obtained with MATLAB can be seen in Fig. 3.9. The results show close computation time of both schemes until 500 UEs, then

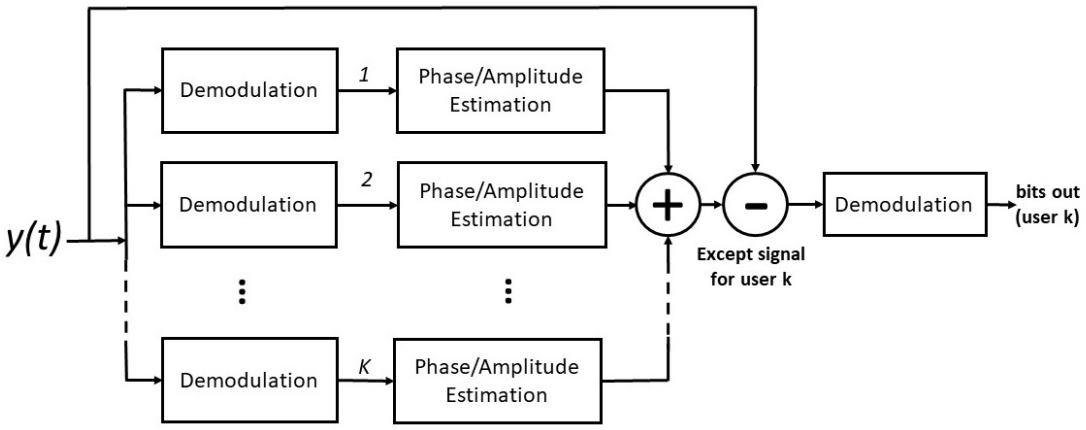

FIGURE 3.8: Block diagram of PIC receiver for NOMA with PIC in the uplink channel.

TABLE 3.3: Simulation Parameters

Parameter Name	**Parameter Value**
Transmit Power (P_T)	23 dBm
Propagation Model	Okumura-Hata
Noise density (N_0)	10^{-17} W/Hz
Frequency	1 GHz
Number of UEs on CPU with MATLAB (K)	500 - 2000
Number of UEs (Java and CUDA)	500 - 2500
Trials	100 000 Times
Bandwidth (W)	50 MHz
Modulation (W)	QPSK and 16-QAM
Software Parameters	
Software CPU	MATLAB (tic-toc)
Software CPU (multi thread)	Java (version 8)
Software GPU (multi thread)	CUDA
Hardware Parameters	
CPU	Intel® Xeon® CPU E5620
Physical Cores	4*2.4 GHz
Memory (RAM)	5 GB 1333 MHz DDR3
GPU	NVIDIA® TITAN Xp
GPU Cores	3840 CUDA Cores
GPU Boost Clocks	1582 MHz

the computation time for SIC gradually increases, whereas the PIC computation time rises moderately. PIC for both modulation schemes have similar curves and 16-QAM version overlaps the QPSK. The substantial difference is recognised for the maximum UEs per cell. The growing difference can be explained by the dependency of decoding among the

UEs for SIC receiver and against the feature of independent decoding of PIC receiver.

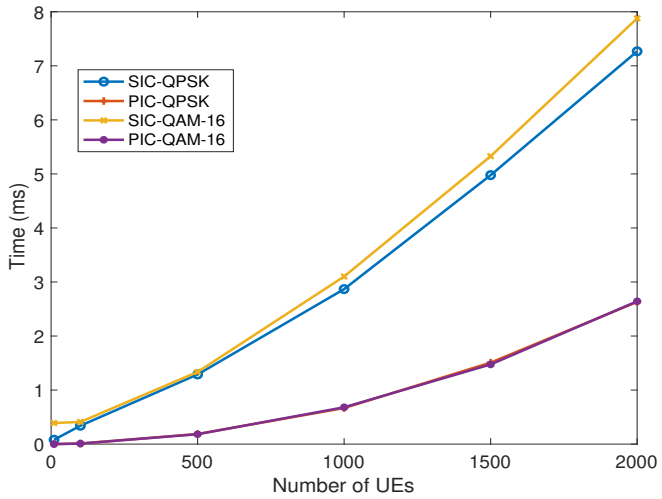

FIGURE 3.9: Computation time of SIC and PIC for QPSK and 16-QAM modulation schemes with CPU.

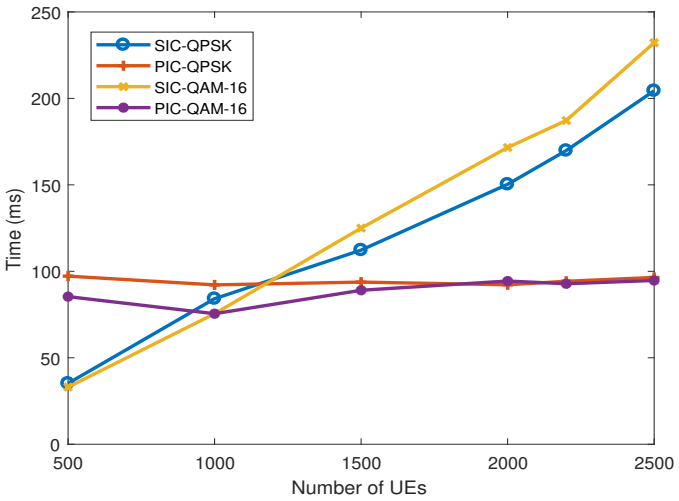

FIGURE 3.10: Computation times of SIC and PIC run with Java version 8 on multiple CPU threads for QPSK and 16-QAM modulation schemes.

Fig. 3.10 shows the execution times of SIC and PIC receivers for 500 to 2500 UEs per cell. ALthough 5G cells are not expected to support this many users, we increased the number of UEs in order to observe the computational behaviors on different platforms. Both receivers were executed on multithreaded CPU with Java programming language, which allows initiating separate threads [126]. The computation time of parallel cancellations with Java retains around 100 ms for any number of UEs. Multiple parallel CPU

threads were initiated per UE and run in parallel. Threads wait for their order in queues and cannot execute 500 or even 2500 operations simultaneously due to a limited number of physical cores. However, the number of simultaneous tasks may be increased to a number of virtual cores in a CPU.

SIC computations with Java took less time than PIC until 1000 UEs and then curve of SIC grew gradually as compared to relatively stable curve of PIC computation time. SIC performs faster due to the acceleration of the processor with less number of UEs. The power of CPU is not spent on parallelization, managing threads and ordering the queues. The power of the CPU is directed only on decoding and cancelling until the needed signal. However, a large number of UEs makes the process too heavy in order to decode the 1000 or more UEs. It may be concluded, that the number of UEs does not affect execution time for PIC due to parallel streams running in multicore CPU.

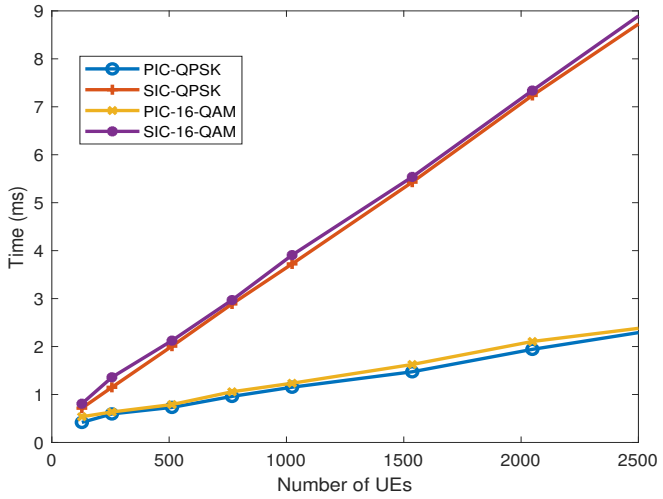

FIGURE 3.11: Computation times of SIC and PIC run with CUDA platform on multiple GPU threads for QPSK and 16-QAM modulation schemes.

Multi-threading may be enhanced with NVIDIA® GPUs by assigning a thread to a particular task. In this case, PIC for each desired signal is implemented on a separate CUDA thread. The CUDA platform is run along with C++ programming language. The results show a slow increase for PIC and stable growth of about 1ms per each 250 UEs for SIC receiver (see Fig. 3.11). For the largest estimated cell size of 2500 UEs it took

about 9ms for SIC and only 2ms for PIC. The behaviour of both modulation schemes is the same.

Although the computation time for both modulation schemes were found close, their error rate performances are different. Fig. 3.12 compares the bit error rate of the closest UE to the base station for QPSK and 16-QAM when $K = 10$ and SIC receiver is used. As expected, lower modulation order offers higher error performance for the same SNR. In SIC receivers, however, increasing SNR further will not improve the error performance since the system is interference limited [29]. In Section 4.2, BER performances of the SIC and PIC receivers are further discussed and compared.

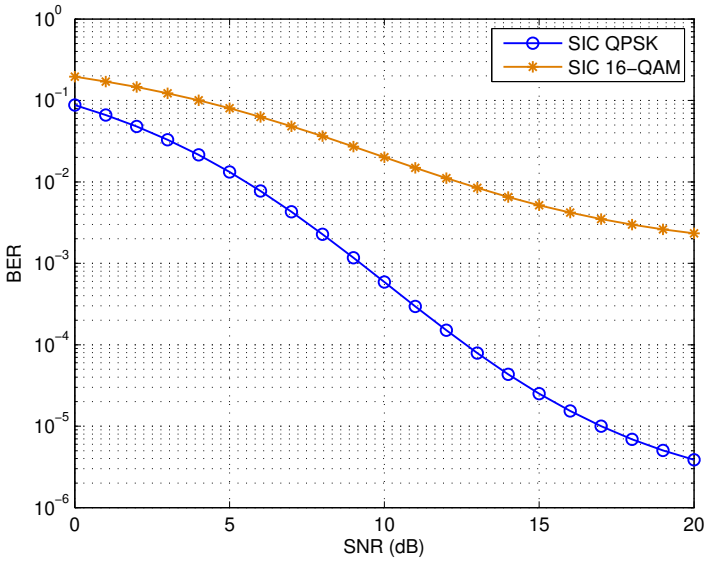

FIGURE 3.12: BER of the closest UE to the BS when $K = 10$.

3.4 Chapter Summary

This section concludes the work described throughout the chapter. In the first part, the thesis discussed optimum power allocation for NOMA with SIC for different fairness constraints. The principle of the algorithm is to aim at the highest cell capacity values within the predefined fairness constraints. This allows the maintenance of the balance of possible data throughout the UEs. The NOMA with SIC capacity with 0.5, 0.7 and

0.9 fairness constraints outperformed OMA capacity with 0.9 fairness constraint. Higher fairness constraint reduced the sum capacity of the network.

The second part considered possible threats upon NOMA with SIC in the downlink channel. The distance between the UE and the BS in NOMA systems defines its decoding order which accordingly relates to the anticipated time for decoding. Unusual time of a particular UE may be reasoned by the fraud and the simulation contains mistakes in the decoding order. The numerical results show the time spent at the UE to decode in the correct order and cases when they attempt to decode other adjacent UEs. The proposed remedy is to complicate the transmission algorithm which may require auxiliary devices.

In the third part, the execution time of the receiver at the NOMA BS with PIC and SIC schemes were compared on different platforms. Decoding of a signal belongs to a UE in SIC depending on other UEs, thus, resulted in gradually increasing time with the number of UEs in the network. PIC receiver time changed moderately due to its parallel manner. Implementation times of these two receivers were evaluated with initiated multiple threads per CPU and GPU cores. CPU implementation is written in Java programming language and GPU software is written with C++ programming language on the CUDA platform. CUDA is a programming platform by NVIDIA® for developing general purpose applications on GPU threads. Performance of CUDA application became the fastest, and the architecture of the PIC scheme came out as less dependent on the number of UEs comparison to SIC.

Chapter 4

GPU Accelerated NOMA

This chapter deals with GPU acceleration in clustered NOMA systems with SIC receiver (Section 4.1) and OFDM-NOMA systems with both SIC and PIC receivers (Section 4.2). Both system models are given for the uplink channel such that the interference cancellation is executed at the BS. The system models findings inferred in the literature are discussed in (Section 2.3.1) to (Section 2.5). Specifically, clusters proposed in the literature limit their models to user pairing, the research project system model considers large independent clusters. Also, SIC receivers dedicated to each cluster run in parallel with the help of GPU accelerators that deem a novel cell architecture. Similarly, OFDM-NOMA systems are discussed with conventional and modified SIC and PIC receivers. Lastly, execution times of the receivers on both CPU and GPU are compared.

4.1 User Clustering in SIC NOMA

4.1.1 Introduction and Related Works

User clustering is an important and major technique that is exploited for enhancing the cell data throughput and the number of UEs supperted within the cell. Beamforming at the BS builds a cluster by pairing the users with the highest channel gain differences and improves the sum capacity of NOMA in [127]. The capacity of UEs with weak channel gain is proportional to the number of UEs in a cell. A large number of users cause inter-user interference due to neglecting the user scheduling which also leads to low capacity.

The study in [128] considers several ways of beamforming to meet QoS for all UEs in the cell. Unfortunately, there is little research done on user clustering. Works in [55], [44] discuss NOMA user-clustering and power allocation for multiple UEs in both downlink and uplink channels. Similarly, authors in [129] apply spatial filtering before the SIC and hoped to remove the inter-beam interference in MIMO-NOMA systems in both uplink and downlink channels. Researchers in [130] propose an algorithm for uplink NOMA with single carrier frequency division multiple access (SC-FDMA) user scheduling. Optimal user grouping was achieved with channel gains of separate UEs. Dynamic behaviour of UEs, errors in channel estimation and time delays in feedback continue to challenge user clustering faces. The study in [131] considers zero-forcing beam design and user clustering with conditions of user fairness, awareness of partial and full channel at the BS in the downlink channel. It also proposes two algorithms upon clustering and optimum power allocation algorithm. The works in [30], [129], [42], [127], [132], [133] consider NOMA with MIMO which matches user clustering conceptually. The size of the clusters varied starting from paired users (2 UEs) and some with 4 UEs and others with several inter and intra-clusters. The research referenced above considered performance evaluations, dynamic power allocation for user clustering, beamforming techniques and complexity analysis. However, there does not appear to be published research which considers NOMA clusters with GPU hardware resided on the BS.

4.1.2 System Model

In this research model, the uplink of a single cell NOMA network with K UEs is considered. The signals are transmitted from each UE to the BS simultaneously using single carrier modulation schemes such as QPSK, 16-QAM or 64-QAM. The propagation paths for each UE differ, and thus, the signals reach at the BS at different power levels [23]. The received signal for K UEs at the BS is represented as

$$y(t) = \sum_{k=1}^{K} \sqrt{P_T} x_k(t) g_k + n(t) \tag{4.1}$$

where g_k is the channel attenuation factor for the link between the BS and each UE_k, $n(t)$ is the AWGN at the BS with zero mean and density of σ_n^2, $x_k(t)$ is the message signal transmitted by the UE k and finally P_T is the total available power of each UE. Considering perfect cancellation, the data rate for the kth UE becomes

$$R_k = \log_2\left(1 + \frac{(P_T)g_k^2}{\sum_{i=k+1}^{K}(P_T)g_i^2 + \sigma_n^2}\right). \tag{4.2}$$

There is a lot of research published addressing the power allocation and some of them were referenced in Section 3.1.1. However, uplink NOMA emphasises channel gain differences rather than adjusting the transmitting powers of UEs. Moreover, since optimum power differentiation is subject to only SIC receivers, here the research considers equal received powers of all UEs for both SIC and PIC receivers that move the research focus to their computation performances, not their reliability.

The number of iterations in the SIC receiver is equal to the number of UEs in the cluster. The drawback of SIC is that where there are more iterations there are more chances to decode incorrectly and carry the error in the next iteration cycles. Moreover, the processing time increases as well [29], [41], which contradicts the latency requirement of the expected 5G systems 5G systems [3], [4], [38]. Maximum likelihood detector searches among every potential binary vector during the demodulation. The computation time of the receiver is also affected by the modulation order. According to Vannithamby and Talwar [134], some crowded cities are expected to serve thousands of active users in 5G cells. To sum up, there is a challenge for SIC receiver to decode large number of UEs and keep a high data rate.

4.1.3 User Clustering in Uplink NOMA

The idea of grouping connected devices into one or more clusters is relatively novel and therefore only a few recent works have been written on this topic. Some of them are reviewed in Section 2.4. The idea of clustering NOMA is recognized as a way of producing

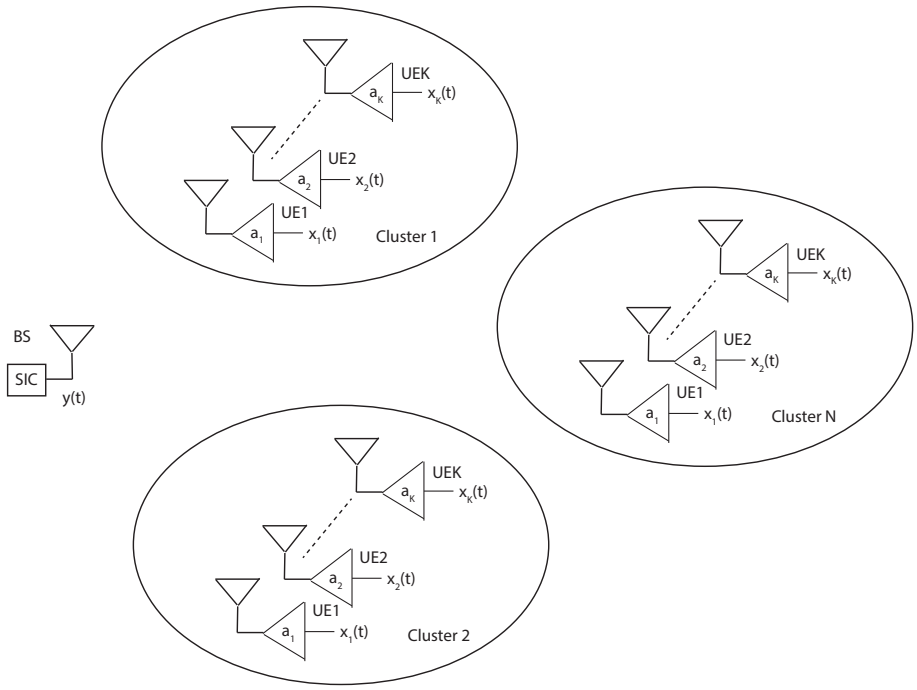

FIGURE 4.1: NOMA with multiple clusters in uplink channel

better data throughput and increasing coverage area if high shadowing losses or blockages occur [55], [134]. According to Ali et al. [55], [135] the system resources are shared per cluster and in some cases, MIMO solutions support multiple transmissions. Technical parameters such as several connected devices per cluster, number of clusters per cell and the way UEs have optimally grouped as well as the distribution of the spectrum and energy are studied in [55], [136], [99].

In the research model, a single NOMA cell in the uplink channel consists of several clusters each having K UEs is considered (Fig. 4.1). Such groupings of the users in clusters make SIC to be processed separately for each cluster in parallel and significantly reduces the workload of the BS. Particularly, BS may initiate and operate multiple SIC receivers for different clusters simultaneously and execute interference cancellations many times faster with its built-in GPU.

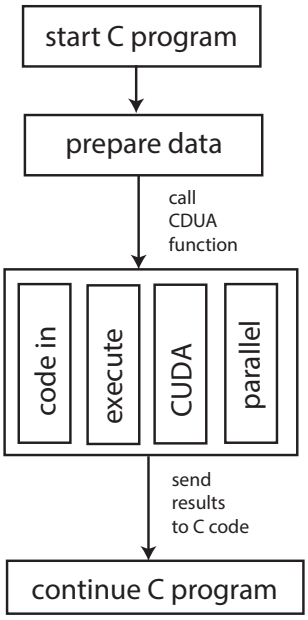

FIGURE 4.2: Steps of running a CUDA C program

4.1.4 GPU Implementation

GPU Based Computing

The software application used in this Chapter was developed in C programming language and CUDA platform by NVIDIA®. Besides C, CUDA is compatible with C++, Python, FORTRAN and MATLAB. Originally, CUDA was working and developing with C and gained popularity among the developers. The program code looks like an ordinary C program with some additional keywords that address CUDA platform. It allows starting of the software on a host (CPU), prepare data (received signal by the BS), call CUDA function to run some part of the code on a device (GPU) and send the prepared data to the CUDA function. The CUDA part initiates the needed amount of threads on GPU, runs the code in parallel and sends the results back to CPU. Then, the C program continues its running with the obtained results (see Fig. 4.2).

Each step of calling a CUDA function carries some preparation procedures. The prepared data has to be copied to the device memory. Also, the number of parallel sub-tasks must be defined. Sub-tasks are executed on device threads. There are thousands of threads on a GPU depending on a model and they are responsible for one of the smallest operation.

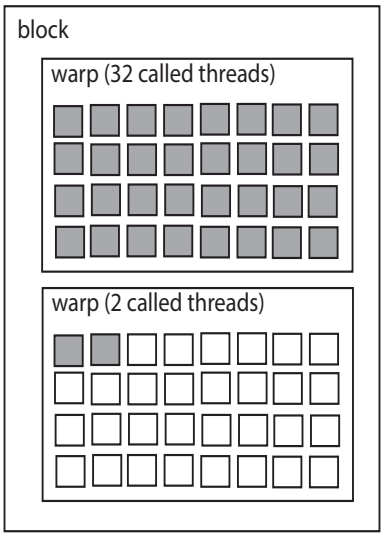

FIGURE 4.3: A scenario of unneeded CUDA threads are called

Hundreds of threads constitute as a CUDA block. Blocks are structured into dimensions. The number of these features are set into CUDA function parameters and the programmer needs to calculate the required number of them. This allows for the avoidance of data and resource leakage. Hardware will group 32 threads into a warp which should be kept in mind during a CUDA function call. For example, if a CUDA function is called with 34 threads, then GPU will initiate two warps one with 32 threads and another with only 2 threads in a new warp with empty 30 threads running (see Fig. 4.3). After finishing the CUDA function the data must be copied back into CPU memory.

Following code has excerpts from CUDA C program. It shows how memory for GPU variable is prepared and processed. It is not a full and continuous working code. The comments are describing the code.

```
/*
define size of the FFT
*/
#define FFT_size                512
#define threads                 32
#define blocks                  4
/*
initiate variable for real part of the received signal on CPU
```

```
and allocate memory on CPU
*/
float *signal_R = (float *)malloc(sizeof(float)* FFT_size);
/*
initiate variable for real part of the received signal for GPU
and allocate memory on GPU
*/
float *d_signal_R;
cudaMalloc((void**)&d_signal_R, FFT_size*sizeof(float));
/*
copy data from CPU to GPU
*/
cudaMemcpy(d_signal_R, signal_R, FFT_size*sizeof(float),
cudaMemcpyHostToDevice);
/*
call CUDA function and send variables as parameters
*/
fillSignalArray << <threads / blocks >> >(d_signal_R);
/*
copy variables back to CPU after CUDA function is finished
*/
cudaMemcpy(&signal_R, d_signal_R, FFT_size*sizeof(float),
cudaMemcpyDeviceToHost);
/*
free allocated space after using the variable
*/
cudaFree(d_signal_R);
```

Copying data from CPU to GPU memory is an actively discussed bottleneck of accelerators. In CUDA, memories are categorized into global memory, texture memory, constant memory, shared memory and registers. Shared memories and registers operate fast. However, they cannot be used for storing variables of large size. Registers are initiated only within a thread, whereas shared memory has a scope of the block. Moreover,

shared memory physically operates on GPU itself, and therefore it was found suitable for the scope of our clustering task. Global memory can store the largest possible size and is, therefore, the slowest one. There are works with more technical details about CUDA memory categories, rather than descriptions [137].

GPU Solution

According to the research scenario is that each cluster has different connected UEs and implements SIC independently. Projection of the proposed solution on the memory of CUDA is illustrated in Fig. 4.4. One cell fits into one CUDA dimension and this research project uses CUDA blocks and CUDA threads that handle clusters. In this experiment, 32 threads per block to fill physical CUDA warp are used. During experiments, the number of called blocks were changed so that the number of threads could be kept as divisible to 32. Algorithm 2 shows SIC which was executed on each thread. Each initiated thread runs the loop K times according to the number of UEs per cluster.

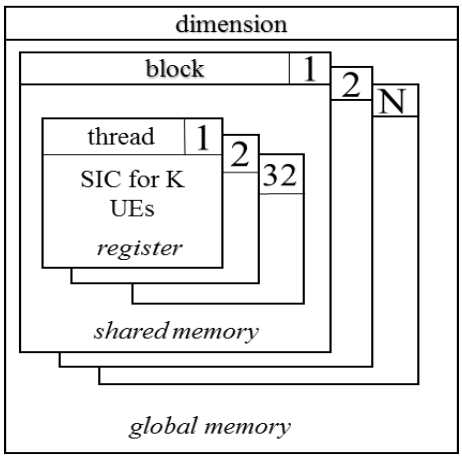

FIGURE 4.4: A SIC for a single cluster projected on CUDA memory

4.1.5 Numerical Results and Discussion

This section shows numerical results obtained by clustering the uplink NOMA cell and implementing SIC on GPU on CUDA platform. There are 10 UEs per cluster and we experimented with 32, 96, 160 and 192 clusters per cell. The research used 32 threads

Algorithm 2: Successive interference cancellation

```
received data: a_k's, g_k's, y(t) for a cluster for each i in {K, .., 1} do
    Equalize the received signal for UE i using a_i and g_i;
    Decode and obtain binary sequence for UE i;
    Modulate the binary sequence to obtain x_i(t);
    Align the phase/amplitude of x_i(t) using a_i and g_i;
    Update y(t) as y(t) = y(t) - x_i(t);
```

per 1, 3, 5 and 6 blocks. Product of blocks and threads multiplied to ten (number of UEs per cluster) gave the total number of UEs per cell. For instance, 3 blocks with 32 threads multiplied to 10 give us 960 UEs in a cell.

TABLE 4.1: Simulation Parameters

Parameter Name	**Parameter Value**
UEs per Cluster	10
Clusters per cell	32, 96, 160 and 192
Transmit Power (P_T)	23 dBm
Propagation Model	COST-Hata
Noise density (N_0)	-174 dBm/Hz
Modulation (W)	QPSK, 16-QAM, 64-QAM

SIC receiver with the above settings was implemented on both GPU and CPU. GPU hardware is NVIDIA Titan Xp with 3840 cores and CPU is a quad-core processor, more details on the experimental setup are given in Table 4.2. The scenario of the experiment considers a large number of uniformly spread UEs in the cell, which is then grouped into clusters. COST Hata model was used to calculate the channel gains g_k for each UE. Transmit power of a UE P_T and noise density σ_n^2 are set as 23 dBm and -174 dBm/Hz, respectively. These features affected the following connection parameters interference among clusters, error propagation, signal-to-noise ratio (SNR) and bit-error-rate (BER). Whereas decoding errors did not prevent it from running all SIC iterations.

Table 4.3 shows the measured execution times for SIC implementation in a cell without clustering on CPU, clustered cell on CPU and GPU hardware with different modulation parameters. The experiment was repeated for different modulation orders and with different number of UEs. In the table, the third and the fourth columns are the results of SIC computation on conventional CPU without and with user clustering, respectively.

TABLE 4.2: Experimental setup

	CPU	GPU
Platform	Intel core i7 (Quad-core)	NVIDIA Titan Xp 3840 Core
Clock rate	2.3 GHz	1582 MHz
Memory	DDR3 2GB*2	DDR3 4GB*3
Language	MATLAB	C & CUDA 9.0

The last column gives this execution time with user clustering using GPU hardware. Proposed GPU solution allowed the research to accomplish the process within 0.2 ms for any number of UEs and all three modulation schemes.

It was concluded from the results that each cluster should contain a scarce number of UEs. The reason was that clustering on CPU requires the sequential implementation of SIC for all UEs in a cell. This led to running SIC for 320 UEs with QPSK modulation in about 0.9 ms and exceeded it up to nine times for 1920 UEs which made CPU impractical for NOMA with SIC. It may be concluded that user clustering and parallel programming could be a solution to envisage a large number of UEs in the upcoming 5th generation of wireless cellular networks.

TABLE 4.3: Comparison of computational times using GPU and CPU in ms.

	Number of UEs	CPU (w/o clustering)	CPU (with clustering)	GPU (with clustering)
QPSK	320	0.851	0.784	0.1495
	960	2.933	2.767	0.1495
	1600	6.594	4.795	0.1505
	1920	9.441	5.346	0.1516
16-QAM	320	1.557	0.981	0.1505
	960	3.388	3.165	0.1516
	1600	6.917	5.002	0.1516
	1920	9.678	5.723	0.1526
64-QAM	320	1.745	1.001	0.1505
	960	3.751	3.382	0.1516
	1600	7.679	6.236	0.1527
	1920	9.783	7.957	0.1536

4.2 PIC and SIC for OFDM-NOMA

4.2.1 Introduction and Related Works

This section discusses the GPU acceleration for OFDM-NOMA based systems. The tasks of the receivers in NOMA cell comprise accurate decoding and interference cancellation. Most of the researches on OFDM-SIC receivers aim at either their error or computation performances [138]. This study similarly attempts to enhance the latter while considering the receiver error performance.

The work in [139] presents a novel solution to cope with the need for a large power difference of OFDM-NOMA in power domain user pairing. The works in [91, 140–143] consider resource allocation and optimisation of several system performances in OFDM-NOMA wireless networks. In most papers, again user-pairing limits the cluster size and rely on the power differences only between two UEs. Otherwise, equal power and geometric programming are used which lead to capacity loss and high computational complexity.

The study in [144] investigates cognitive OFDM-NOMA with spectrum utilisation. The work maximises the sum capacity with a two UEs per subcarrier scenario and further expands the research to build a power allocation algorithm. Similarly, [45] studies the power allocation for NOMA and proposes two methods to maximise sum rates comparing this to an optimal one. The references [145] and [146] consider the issues which occur when a receiver codes several UEs. Firstly, error propagation and secondly time delay until all the UEs are decoded prevent SIC to be implemented in the receivers. User grouping and power allocation reduce the number of UEs in a queue and result in higher sum capacity. The study in [147] has an optimistic view on receivers of upcoming mobile systems and expects processors of the BSs to make the NOMA based advanced receivers possible in real life.

The studies in [138] and [148] emphasise that PIC receivers need strong hardware to cope with the tasks in parallel and found the GPUs impractical and costly. However, GPU devices today have become massively available and substitute CPUs in complex tasks

where parallel implementation is possible [149]. Most of the works in the literature put effort on the optimisations of the physical layer parameters and optimistically rely on the powerful receivers of the future. From the research carried out, there appears to be no research that proposes the architecture of the receivers on GPU devices yet. Moreover, there are no studies on the parallel architecture of the receivers and the way SIC and PIC could be implemented on GPU hardware with the CUDA programming model or any other.

4.2.2 System Model

We consider an uplink of a single cell NOMA-OFDM network with K UEs. All N sub-carriers of the OFDM symbol are occupied by every UE of the cell. Then, at the BS, UE signals are differentiated by one of the PIC or SIC interference cancellation techniques. We assume that the channel gains are in the order of $|\alpha_1|^2 > |\alpha_2|^2 > ... > |\alpha_K|^2$, where $|\alpha_K|^2$ is the channel gain of the farthest user. The signal received by the BS will be the summation of the individual signals transmitted by each UE and can be formulated as

$$y(t) = \sum_{k=1}^{K} \sqrt{P} s_k(t) \alpha_k + n(t) \tag{4.3}$$

where P is transmission power for each UE, $s_k(t)$ stands for the sent OFDM symbol by the k^{th} UE, α_k is the channel attenuation gain between the k^{th} UE and base station and additive white Gaussian noise with σ_n^2 variance is $n(t)$. The discrete time transmitted OFDM symbol by the UE k can be written as

$$s_k[m] = \frac{1}{\sqrt{N}} \sum_{n=0}^{N-1} S_k[n] e^{j2\pi mk/N} \tag{4.4}$$

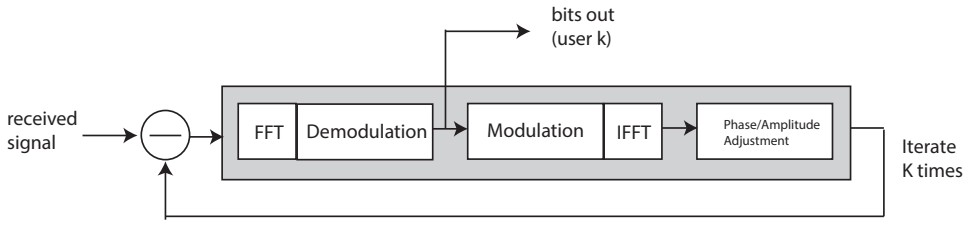

FIGURE 4.5: Iteration step of a successive interference cancellation (SIC) with orthogonal frequency division multiplexing (OFDM).

where $S_k[n]$ is either PSK or QAM modulated complex symbol at the subcarrier n and N is the OFDM size. In continuous time, OFDM waveform is written as

$$s_k(t) = \sum_{m=0}^{N-1} s_k[m]p(t - mT_s) \tag{4.5}$$

where T_s is the baud rate and $p(t)$ is the pulse-shaping filter. BS receives the summation of all the transmitted OFDM signals as in (4.3). Then to decode the individual messages, it removes interference either with SIC or with PIC scheme. This system model involves OFDM modulation and demodulation in both interference cancellation schemes. The fast Fourier transform (FFT) and inverse fast Fourier transform (IFFT) procedures deal with complex symbols on each subcarrier. These complex symbols are mapped to or from bit sequences by modulation or demodulation blocks. The interference cancellations with FFT and iFFT operations are illustrated in Fig. 4.5 for SIC and in Fig. 4.6 for PIC. In CPU implementation, i.e., in MATLAB, fft and ifft operations are implemented with parallel programming involving multiple CPU cores, however, other operations in the remaining part of the SIC or PIC receivers such as modulation, demodulation, loops etc. are still sequential.

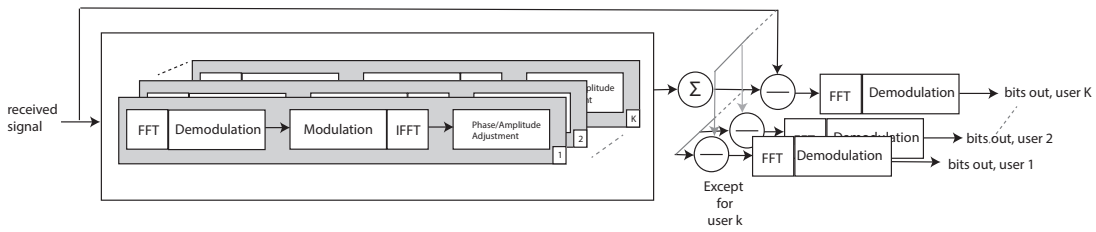

FIGURE 4.6: Parallel interference cancellation (PIC) with orthogonal frequency division multiplexing (OFDM) for UE K

4.2.3 OFDM based SIC and PIC in NOMA

The SIC receiver decodes the signals one after the other by iterating the cancellation process. The sequence starts from the UE with the strongest signal and ends with UE with the weakest signal. It relates to the invidual user's distance to the BS, as it closer to the BS its signal gets stronger. In each iteration, while the signal is decoded, other signals of the rest UEs are treated as interference. BS may initiate additional operations to prevent errors in the obtained bit sequences, like re-transmission and regenerate the transmitted OFDM signal. Then, amplitude and phase of the OFDM signal are adjusted as the signal passed through the channel between the UE and the BS. Finally, the regenerated signal is subtracted from the received signal and the steps described above are iterated for all UEs that are participating in the transmission. The SIC with OFDM is illustrated in Fig. 4.5. In case of ideal interference cancellation and when regeneration process had conducted successful phase/amplitude adjustment, the time domain OFDM signal for k^{th} UE at the kth iteration can be written as [150]

$$\hat{s}_k(t) = \sqrt{P}\alpha_i s_k(t) + \sum_{i=k+1}^{K} \sqrt{P}\alpha_i s_i(t) + n(t). \tag{4.6}$$

The BER and SINR per subcarrier are considered as the average values due to the frequency flat channel conditions of the considered channels. The SNR and the BER for binary transmission are

$$\text{SINR}_k = \frac{P|\alpha_k|^2}{\sigma_n^2 + \sum_{i=k+1}^{K} P|\alpha_i|^2} \tag{4.7}$$

and

$$\text{BER}_k = Q(\sqrt{2\text{SINR}_k}) \tag{4.8}$$

respectively. The $Q(.)$ in (4.8) is the distribution function of the standard normal distribution [151].

The PIC receiver works differently than SIC in many ways and its working structure can be seen in Fig. 4.6. The process can be logically separated into two stages. In the first stage, all the received signals of other than particular UE are decoded, regenerated,

summed and then subtracted from the received signal. In the second stage, OFDM demodulation is performed for the resultant signal. This processes is done for each UE and are independent of each other. Thus may be implemented in parallel for all UEs. The time domain OFDM signal $\hat{s}_k(t)$ for particular k^{th} UE is represented as

$$\hat{s}_k(t) = \sum_{i=1}^{K} \sqrt{P}\alpha_i s_i(t) - \sum_{i=1, i\neq k}^{K} \sqrt{P}\alpha_i \hat{s}_i(t) + n(t). \tag{4.9}$$

For PIC, BER of binary transmission for the k^{th} UE per subcarrier is written as [152], [153], [154]:

$$\begin{aligned} BER_k &= Q\left\{ \left[\frac{1}{2(P_k/\sigma_n^2)} \left(\frac{1-(\frac{K-1}{3})^{s+1}}{1-(\frac{K-1}{3})} \right) + \frac{1}{3^{s+1}} \right. \right. \\ &\quad \left. \left. \times \left(\frac{(K-1)^{s-1} - (-1)^{s+1}}{K} \left(\frac{\sum_{i=1, i\neq k}^{K} P_i}{P_k} + 1 \right) + (-1)^{s+1} \right) \right]^{-1/2} \right\} \end{aligned} \tag{4.10}$$

here s stands for the number of stages in the PIC receiver. As discussed above, in this case, there are two stages. CDMA schemes may have multi-stage receivers with unit processing gain [153] which lead to this akin equation (4.10). In CDMA systems, the difference of origin equation is in distinguishing UE signals by assigned signature codes. In these comparisons, code domain NOMA variant is considered for PIC to its analogous reliability of interference cancellation schemes, their computational expenses and performance similarity [155], [156]. The derivation is built upon decoding errors which occur during the first stage.

4.2.4 GPU Implementation

Next, GPU implementation of SIC and PIC techniques on CUDA is described. Computational speeds of both receivers under various scenarios are compared on two different machines: one for CPU results and the other for GPU results. The machine for parallel computation is equipped with Intel Xeon CPU E560 @ 2.40GHz, 9GB RAM with DDR3 1333 MHz and built-in NVIDIA TITAN Xp graphics card with 12 GB memory.

The GPU has 3840 CUDA cores and 1582 MHz clock rates. Software part had C++ object-oriented programming language and CUDA platform with Nvidia CUDA Compiler (NVCC) which was updated to version 9.2. As for the machine that was used for conventional serial programming, it has Intel core i7 CPU with 4 physical cores and 2.3 GHz frequency, 16 GB DDR3L RAM with 1600 MHz frequency. The software application was implemented in MATLAB R2018 IDE.

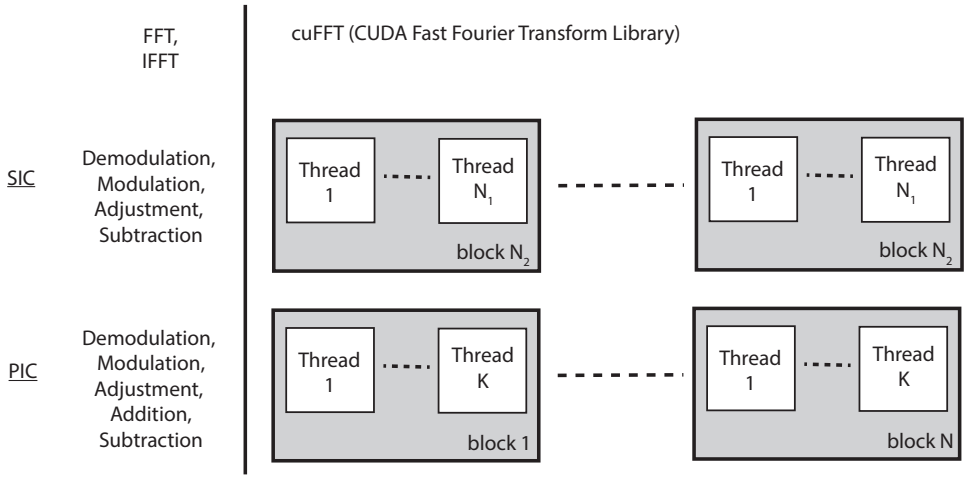

FIGURE 4.7: CPU and CUDA functions of SIC and PIC with blocks and threads on CUDA.

Normally, functions of a program run on CPU. A kernel is a CUDA function written by a developer, which runs on a GPU device. Fig. 4.7 shows functions, kernels and library functions which are involved in NOMA SIC and PIC schemes. Fig. 4.7 also demonstrates the architecture of CUDA threads and blocks for each of those schemes. CUDA platform has a special cuFFT library for running FFT functions with forward and inverse parameters. It has a $\mathcal{O}(nlogn)$ computation time complexity. NVIDIA Titan Xp graphic card is limited to fit CUDA 1024 threads in a block. For our kernels, we called one CUDA thread per subcarrier.

As it can be seen in Fig. 4.5, SIC scheme has FFT, IFFT, demodulation, modulation, IFFT and phase/amplitude adjustment tasks. Special cuFFT library is included to implement FFT and IFFT tasks [157]. Kernel versions of modulation, demodulation and subtraction the decoded signal from the received signal were developed. Even though

SIC execution is sequential in nature, OFDM computations still can be called in parallel per UE according to divide and conquer algorithm [158]. Each kernel runs on N_1 blocks with N_2 threads per block. $N_1 \times N_2 = N$ gives us the size of a grid equal to the number of subcarriers. N tasks run in parallel for only one UE in a cell and then SIC receiver iterates these steps for all UEs.

In contrast, the PIC receiver neatly suits to CUDA architecture for running on a GPU. One thread is initiated for executing a task of a subcarrier. One block has K threads for each UE. The number of blocks equals to N, which is the size of FFT. Finally, the size of a grid becomes as $K \times N$. CUDA architecture of PIC is illustrated at the bottom of Fig. 4.7. It can be recognized that a CUDA kernel for parallel addition task was developed which was not needed in SIC.

4.2.5 Numerical Results and Discussion

This section presents the numerical results obtained by PIC and SIC receivers for NOMA-OFDM systems on GPU. Firstly, BER performances and then computation times of both receivers on CPU and GPU are compared. Numerical results are obtained with the different number of UEs with an active connection to the network. UEs' channel gains are set in a descending order with 2 dB difference i.e., $10\log_{10}(\frac{P|\alpha_i|^2}{P|\alpha_{i+1}|^2}) = 2$ dB, like $|\alpha_1|^2 > |\alpha_2|^2 > ... > |\alpha_K|^2$. The received power from the closest UE is $P|\alpha_1|^2$ that is set at -90 dBm power. Some of the simulation parameters are listed in Table (4.4). In the analysis, the UEs are stationary and the channel attenuation is assumed to be the same for all subcarriers of the OFDM symbol.

TABLE 4.4: Simulation Parameters

Parameter Name	**Parameter Value**
Number of UEs	50 - 350
Power from the closest UE ($P^2_{\alpha_1}$)	-90 dBm
FFT size	2048 and 4096

Fig. 4.8 illustrates BER of the first UE with 5 to 50 UEs in the network for SIC and PIC receivers. SNR for the first UE is written as SNR = $P|\alpha_1|^2/\sigma_n^2$ and is equal to 15

dB and 20 dB with σ_n^2 of -105 dBm and -110 dBm, respectively. According to (4.7) for SIC, interference of other UEs restricts the performance. This was depicted in Fig. 4.8, where BER of the first UE and when a large number of UEs exist in the network. This does not depend much on its SNR. (4.10) is used for obtaining BER values of PIC receiver. In contrast to SIC, the rise of SNR affects the BER performance for large number of UEs.

In Fig. 4.9, we plot the BER of SIC and PIC receivers with respect to the SNR of the first UE using (4.7) and (4.10). The noise power is set at -105 dBm/Hz and the SNR differences between the adjacent users is taken as 10 dB, i.e., $10\log_{10}(\frac{P|\alpha_i|^2}{P|\alpha_{i+1}|^2}) = 10$ dB. The BER performance of the SIC receiver cannot be improved further even if the SNR is increased. This is expected since the SNR at the SIC receiver will be dominated by the interference term for small σ_n^2 (see (4.7)). The BER performance of the PIC receiver, on the other hand, improves as the SNR increases. Here, it should be emphasized once again that the BER of PIC receiver in (4.10) is given for code domain multiplexing with unit processing gain.

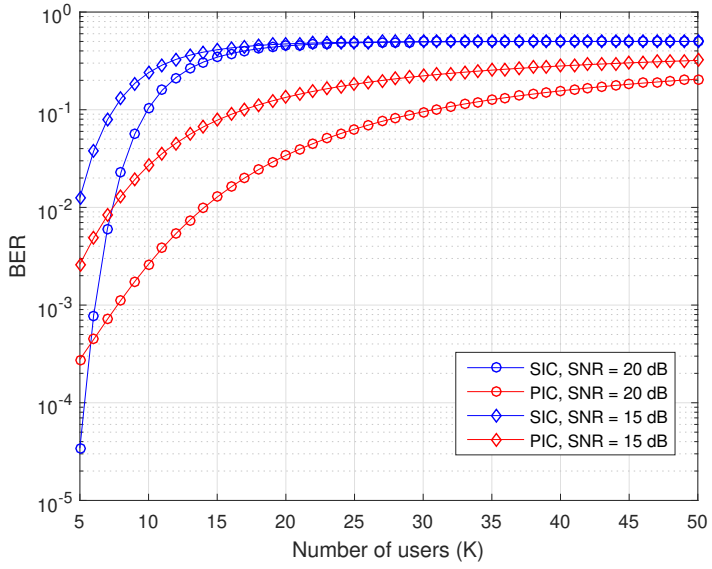

FIGURE 4.8: SIC and PIC bit error rate (BER) for different number of UEs.

The performance of SIC and PIC receivers with OFDM are compared on GPU with CUDA platform and CPU with MATLAB IDE. Results are summarized in Table 4.5 for

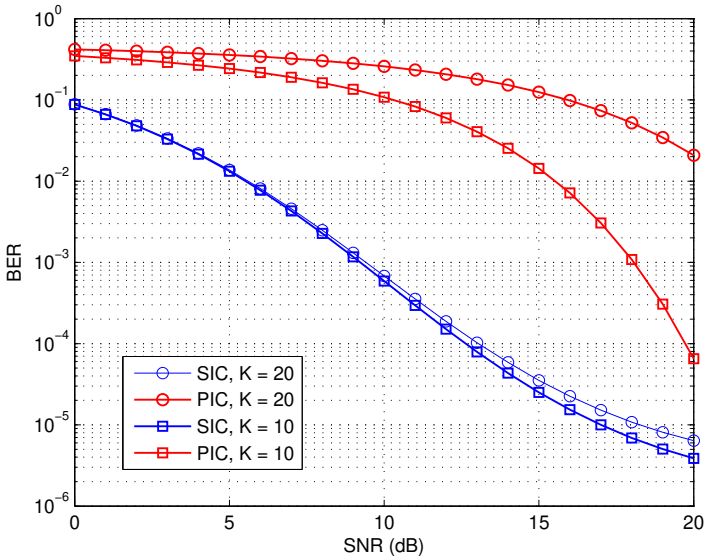

FIGURE 4.9: SIC and PIC bit error rate (BER) for different number of UEs.

OFDM size equals to 2048. Table 4.6 for OFDM size equal to 4096. Maximum likelihood decoding and QPSK are implemented on each subcarrier [159–161]. This means that the interaction time between CPU and GPU during calling functions, copying data between the host and device was not taken into account. MATLAB functions include parallelisation with multiple CPU cores for FFT and IFFT computation functions. The rest of the code attributes like loops are not parallelised in the MATLAB. NOMA cells are expected to only serve about 50 UEs in 5G [155]. However, a far larger number of UEs may fit using MIMO, grouping or clustering UEs [162], [163].The computation results considered from 50 up to 350 UEs.

Before discussing the results, it should be clarified that the PIC receiver has more tasks to execute than SIC. This can be seen from Fig. 4.6. PIC has summation, FFT computation and demodulation steps at the end of the process comparing to SIC (Fig. 4.5). However, the nature of PIC assumes that most of it can run in parallel. Another, detail that should be taken into account is that CPU has fewer stronger processors with 2.3 GHz compared to approximately four thousands of 1.5 GHz GPU cores. Considering those facts, it is clear why SIC outperformed PIC on CPU in both variants of FFT size.

It took about 44 ms to decode 50 UEs and reached 315 ms to decode 350 UEs when

TABLE 4.5: Computation times for SIC and PIC on CPU (MATLAB) and on GPU (CUDA), FFT = 2048.

Number of Users	**CPU**		**GPU**	
	SIC	**PIC**	**SIC**	**PIC**
50	27.70	44.01	68.69	1.97
100	45.49	84.90	138.39	2.08
150	68.22	130.17	208.28	2.19
200	91.05	173.69	281.20	2.21
250	113.73	223.56	353.08	2.23
300	135.94	269.35	426.68	2.26
350	157.98	315.15	500.45	2.28

TABLE 4.6: Computation times for SIC and PIC on CPU (MATLAB) and on GPU (CUDA), FFT = 4096.

Number of Users	**CPU**		**GPU**	
	SIC	**PIC**	**SIC**	**PIC**
50	46.65	87.83	69.22	2.54
100	93.77	172.86	141.49	3.22
150	142.67	263.28	213.03	4.13
200	188.77	349.27	285.72	4.81
250	236.25	442.43	360.51	5.52
300	279.16	530.04	439.31	6.00
350	330.65	616.26	514.95	6.88

FFT size was set as 2048. These numbers doubled and reached an impractical time for decoding the signals when FFT was set to 4096. It took about 88 ms to decode 50 UEs and about 616 ms for obtaining messages of 350 UEs. As became evident from the literature, NOMA with SIC receiver was the most common proposed solution for 5G. Its' results reach about 27.7 ms for 50 UEs and about 160 ms for 350 UEs in a cell with FFT = 2048. Messages of 50 UEs were decoded in about 47 ms and reached 330 ms with more complex FFT size.

Our offered solution of decoding with GPU lead to substantially slower results for NOMA with SIC. It took about 69 ms and 500 ms for 50 and 350 UEs respectively, regardless of the FFT size. The reason is in the iterative repetition of the tasks in a device that possesses more, yet less powerful cores. In contrast, PIC receiver that matches the

device architecture spent about 2 ms to decode any number of UEs with FFT = 2048. When FFT was set to 4096, time to decode 50 UEs was measured as 2.54 ms and reached less than 7 ms for decoding 350 UEs. The experiment revealed that PIC is faster than SIC 75 and 220 times with FFT size equal to 2048 and 4096 respectively. As was explained earlier, the SIC receiver carries data dependency among the UEs, which requires to decode messages one after another for each UE on both devices CPU and GPU.

4.3 Chapter Summary

This chapter demonstrated solutions for NOMA with SIC and PIC receivers in the uplink channel. The first solution was based on the user clustering and allowed to work around the obstacle of user dependency in SIC. The second solution proposed receivers with OFDM computations. The computation times of both cases were compared with the results conventionally obtained with CPU offered in the literature. The research is focused on breaking through the limits of computation time. It may be continued via involving processes which take place in a real life wireless communication. For instance, error correcting or channel estimation. User clustering was implemented on CPU and GPU. These results summarised with SIC receivers without user clustering which also ran on CPU and GPU. NOMA with SIC used user clustering and ran on the GPU allowed to achieve 52 times faster results than CPU. User clustering technique may be applied in the cellular networks to implement parallel programming and boost the system performance.

GPU as hardware helps by providing threads for the larger number of UEs in a cell. Next, the CUDA platform on GPU device was proposed for NOMA with OFDM cellular networks with SIC and PIC receivers. Both types of the receivers were illustrated and their architecture for CUDA platform is designed. SIC and PIC receivers on CPU and GPU devices are summarized with FFT size set to 2048 and 4096. More UEs in a cell resulted in a greater difference between the times obtained with the CPU to with the GPU device. Also, in a scenario with 350 UEs in a cell, PIC receiver was found as 75 times

and 220 times faster than the SIC if both of the receivers ran on GPU with different FFT sizes.

Looking forward to the 5G, complicated operations are expected for baseband processing. Moreover, advances such as MIMO and interference cancellations load heavy processes on the base stations. Results obtained in this research has shown that GPU may run decoding for optimal 50 UEs approximately in 3 ms. This allows involving other possible tasks to meet latency requirements.

Several major obstacles prevent from exploiting the solutions in a downlink channel. Firstly, in order to implement SIC, each UE must know CSI of other UEs in a network. Another reason lies in mobile GPUs which are not ready for general purpose computing. The largest power consumption of mobile power are GPU and communication subsystem [164]. Prior loading communication related tasks on GPU, high temperature and power consumption issues must be solved.

Chapter 5

Artificial Intelligence in NOMA

This chapter deals with NOMA enhancement with machine learning (ML) and deep learning (DL) algorithms to obtain power allocation coefficients. These algorithms are proposed to challenge the complexity in finding the optimum coefficients by exhaustive search. Firstly, a set of power allocation coefficients are obtained for different realisations of UE locations and calculate the peak data rates for each realisation. Then, ML and DL algorithms are applied to predict power allocation coefficients. First, the normal equation of linear regression is considered. Then a deep neural network (DNN) is applied to the same problem. Finally, the execution times of the two methods are presented and a comparison of these is offered with an exhaustive search.

5.1 Introduction and Related Works

The term 'AI' comprises of large ML and DL concepts which have recently been practically exploited in different fields such as medicine, defence, computational finance, weather and climate etc. [165–167]. Solutions obtained by AI often turn out to be more accurate than a human solution [82], [168].

In NOMA, there is published research addressing several AI solutions. For instance, DL to form the channel state information for NOMA system via offline learning from the environment is studied in [169]. ML aided wireless communication architectures improve the system performance via incorporating data [168, 170–172]. [173] and [174]

address ML to group UEs into clusters in mm-wave NOMA systems. DL-aided resource allocation rules in wireless cellular networks were investigated in [175], [176].

Optimal resource allocation is important in NOMA systems not only to ensure there is no waste of the available resources but more importantly it enables the system to work properly. The works in [90, 92, 95, 177–181] propose optimal, suboptimal and heuristic power allocation and channel assignment solutions for NOMA networks. The specific nature of UE channels require non-linear treatment rather than ordinary ones which leads to an overload of computation resources. The power allocation problem is widely researched and mostly corroborated with the sum capacity and UE fairness results [181], [182], [183]. Another work in [99] proposes a power allocation algorithm for OFDM-NOMA with complexity $\mathcal{O}(n^3)$ where n is the number of subcarriers. The work in [184] consider the suboptimal approach founded on the non-linear combinatoric optimization which expects computational complexity that is too high. Furthermore, the works in [57], [108] incorporate the perspectives of cognitive radio in the design of NOMA networks and later expanded for optimum power allocation in [185].

In summary, together with the earlier discussions in (Section 3.1.1), there remains an unmet need for efficient power allocation mechanism in NOMA networks. We list the recent and ongoing contributions in the literature for AI-assisted NOMA networks in Table 5.1. AI aided resource allocation literature is scarce and not conclusive yet. In this contribution, the use of normal equation in ML and deep neural network in DL for the power allocation problem is demonstrated and an evaluation of their execution times is presented.

TABLE 5.1: Summary of technical contributions on AI NOMA

Year	Author(s)	Technical contribution
2016	S. Ali *et al.* [63]	NOMA with K-means clustering outperformed random clustering and MIMO-NOMA clustering based in NOMA proposed and all three ML algorithms in OMA.
2018	R. Amiri *et al.* [186]	The ML-based algorithm with 2^{15} computation complexity attained a high capacity for UEs connected to both additional femto BSs and for macro BSs in HetNets. The brute force algorithm has 2^{75} complexity. Fairness was verified by Jains' index [101].

2018	M. Liu *et al.* [187]	NOMA with imperfect SIC with user pairing for IoT was compared with OFDMA in terms of energy efficiency, sum capacity and total power consumption. AI aided resource allocation. Resource allocation via deep learning method with two hidden layers performed with less complexity for six different conditions of NOMA scheme.
2018	J. Cui *et al.* [173]	Cluster the cell with K-means based ML algorithm. Optimal power allocation within a cluster focus on maximizing the sum data rate and keep a minimum data rate. The simulation results demonstrate that 2 clusters with 8 users optimally keep the distance between the clusters and gradually increase the sum data rate.
2019	F. Jameel *et al.* [188]	The user equipment with strong channel conditions relays message from the BS to a user with poor conditions within nodes. However, eavesdropper which is located in between attempts to decode the signal. The research challenges to predict marginal value of power strength to maximize the secrecy rate with DL.
2019	Y. Sun *et al.* [189]	Obtaining SIC ordering via deep learning alternatively to conventional NOMA with imperfect SIC. Authors compute weights of order for each UE, obtain power allocation and recover order of decoding. Then verify the sophisticated model accuracy, mean squared error, data rate, queue delay and utility results with those of ordered based on channel gain.
2019, 2020	M. Liu *et al.* [185], Y. Zhang *et al.* [190], F. Hussain *et al.* [191]	ML and DL and deep reinforcement learning were exploited for resource allocation in NOMA uplink and NOMA with IoT in downlink.
2020	W. Kim *et al.* [192]	DL was applied to detect active UEs within grant free NOMA systems in the uplink channel. The role of DL is to identify the rule of the received signal for massive machine communications.

5.2 System Model

The system model considered in this Chapter is a special case of the system model presented in Section 3.1. UEs are distributed at varying distances to the BS, but they all lie on one side of it (see Fig. 5.1). The BS modulates the messages of each UE using a single carrier modulation scheme such as QPSK or QAM and then scales and adds up the modulated waves. The SIC receiver at the UE demodulates the received signal starting from the one with the highest power and follows in decreasing order. In each iteration, the demodulated bit streams are used to regenerate the modulated waveform to be subtracted from the received signal. These steps require accuracy of power weights, channel

gain and phase information of the individual signals. The receiver stops iterating when it obtains its message [43]. The data throughput of a SIC receiver in NOMA is [118]

$$R_k = W \log_2 \left(1 + \frac{\alpha_k P_T g_k^2}{N_0 W + \sum_{i=1}^{k-1} \alpha_i P_T g_i^2} \right) \tag{5.1}$$

where N_0 is the noise density (W/Hz), W is the transmission bandwidth (Hz), P_T is the total BS transmission power per signal, g_k is the channel attenuation coefficient for UE k, α_k defines power allocation coefficient for UE k, such that $\sum_{k=1}^{K} \alpha_k = 1$.

It is clear from (5.1) that the data rate and the signal-to-noise ratio of UE k are proportional to power ratio α_k. Earlier in (Section 3.1.1) an exhaustive search method was implemented to find the optimum power allocation coefficients for K users. The method works along with fairness constraints which aim at both: firstly, the highest possible sum capacity and secondly, as equal as possible user data rates. However, the method has serious drawbacks in computation, due to waxing possible power coefficients as the number of UEs in a network raises. Sparkled by the prospects of AI, we have come up with an avenue that outputs power weights via bypassing complicated computations. We first implement basic linear regression, in particular normal equation, and then corroborate with a deep neural network (DNN) model. Input data of AI methods confine with the distances of the UEs to the BS. In other words, features become $[d_1\ d_2\ ...\ d_K]$.

5.3 Data Preparation

There are 11 scenarios each having different number of UEs from 2 to 12 in the network. The distances are randomly selected within [25, 300] metres and are divisible by 25. For example, a scenario with K = 2 UEs may have distances [75, 125] or [300, 175] etc. For each scenario, 682 completely different feature sets as input data are generated. Then, exhaustive search method was run to compute the optimum power allocation coefficients for each feature. These optimum allocation coefficients then became an input data for the loss function. The normal equation used all these 682 examples as input data set at

once. Whereas for the DNN, 618 observations are allocated for training and the rest 64 for testing the algorithm.

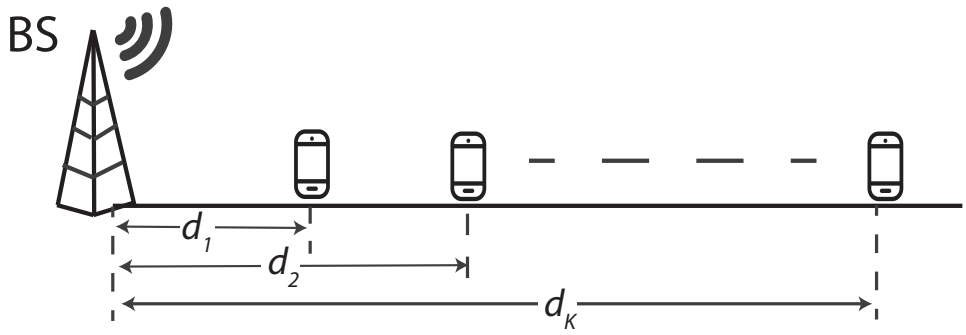

FIGURE 5.1: NOMA system in downlink with K users along one beam.

5.4 AI Implementation

The normal equation in (5.2) computes a closed form solution. X is the data matrix and y is the vector of power allocation coefficients. (5.2) is implemented for each UE in every scenario, changing the values of y from earlier obtained power allocation coefficients of a specific UE. For instance, in a scenario with 2 UEs and a distance vector of [25 75] meters, the optimum power allocation coefficients from the exhaustive search are [0.08 0.92]. Then, normal equation is executed twice. In the first run, X =[25 75] and $y = 0.08$ and in the second run X = [25 75] and $y = 0.92$.

$$\theta = (X^T X)^{-1}.(X^T y). \tag{5.2}$$

The DNN model is illustrated in Fig. 5.2. In the experiment, the parameters were diligently gleaned and the parameters which fostered the best accuracy were selected. As a result, number of users became the size of an input layer, then two hidden layers were added with 4 and 2 neurons. As a result, the number of users became the size of an input layer, then two hidden layers were added with 4 and 2 neurons. As for the activation functions, the rectified linear unit (Relu) was selected for hidden layers and Sigmoid was selected for the output layer. The output value of the model was considered as a power

allocation coefficient of particular UE. This model was called K times for each scenario and power allocation coefficients for all UEs are obtained. However, the sum of the predicted coefficients may have exceeded 1 in some scenarios. Thus, the total power is a constraint to reach utmost 1.

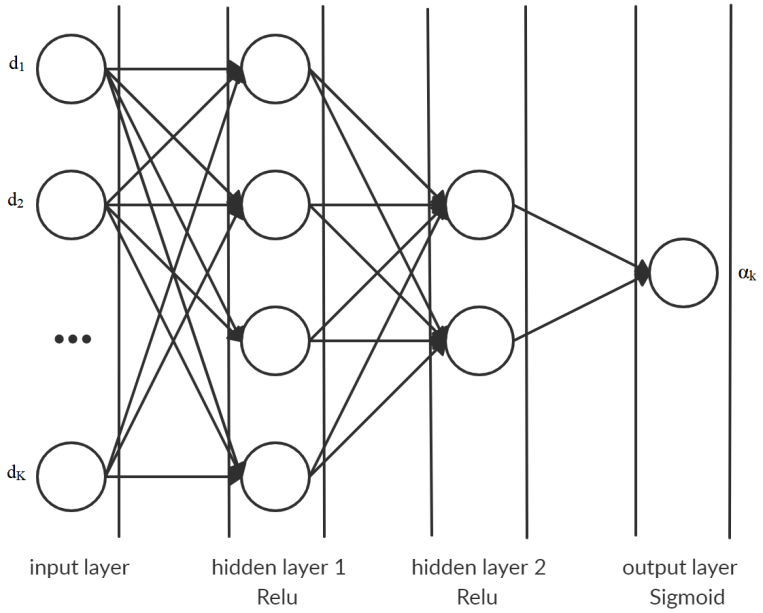

FIGURE 5.2: Deep neural network (DNN) framework to predict α_k.

Hardware parameters of this experiments are as following: Intel(R) Core(TM) i9-7900X CPU @3.30GHz, DDR4 2666 MHz 16 GB Memory RAM. Windows 10 operating system with MATLAB 2019b IDE were used for running exhaustive search and the normal equation, Python 3.7 was used for exploiting deep learning. DNN was implemented using plain numpy libraries without any frameworks such as TensorFlow, Keras, PyTorch, PyCUDA etc.

5.5 Computational Complexity Analysis

This section shows quantity complexity of brute force, normal equation and deep learning algorithms to implement optimum power allocation. We consider FLOP as a real floating point operation [193] and N is an input size. So the complexity of algorithm is number of the FLOPs. Firstly, following rules upon arithmetic operations are considered:

1. 1 FLOP for multiplication, division or addition of 1 real number.

2. N^3 FLOPs Multiplication of two square matrices [194].

5.5.1 Complexity of Brute-Force Algorithm

The Algorithm (1) consists of setting the fairness index, calculating the capacity - (5.1), calculating the fairness index - (3.6), comparing the fairness index and setting values in capacity matrix. Next finding maximum capacity. There are 1 addition, 1 division and 4 + (3×K) multiplications in the (5.1). So the complexity may be written as (3×K)+6. Whereas, in (3.6) there are $2K-2$ summations, $K-1$ multiplications and 1 division, so it can be written as $3K$. Finding maximum part has 1 step of calculating the capacity, which is $(3\times K)+6$, Comparing the calculated capacity and setting a new capacity. so that, finding maximum may be considered as (3×K)+4. Also, (5.1) and (3.6) are iterated throughout the size of the matrix with possible power allocation coefficients and throughout the all non zero capacity values respectively. Finally, the brute force algorithm, may be approximated as 1 + ((3×K)+6) ×25337 + 3K× 6334. We approximated the size of the non-zero capacity array as a quarter of the matrix with power allocation coefficients.

5.5.2 Complexity of Normal Equation

The equation (5.2) has a complexity of O(C^3), where C is the number of features. In our case it is the number of UEs in the network.

5.5.3 Complexity of Deep Learning Model

Our DNN has 1 input layer, 1 hidden layer with 4 inputs then 1 hidden layer with 2 inputs and finally 1 output layer. The size of the input layer becomes as the number of UEs in a network (N). The number of epochs is n, the number of training examples is t. The computational complexity to obtain optimum power allocation for one training example may be written as $O(n\times t\times((4\times N)+(4\times 2)+2))$. Here, we expect one real value as

an output for power allocation coefficient. Then we run this for every UE in the network. Therefore, we multiply the complexity by N. So finally, the deep learning model for obtaining optimum power allocation with N features and t examples may be written as $O(N(n \times t \times (4N+8+2)))$.

To sum up, complexities of the brute force algorithm and deep learning are linear. However, it can be clearly seen that the number of iterations of the brute force algorithm make it times slower than machine learning and deep learning algorithms.

5.6 Numerical Results and Discussion

This section presents the results obtained with the normal equation and with DNN. After running the normal equation for the total data set, θ parameter is obtained. Similarly, after training DNN, the relevant parameter set is obtained. These parameters of both models allow the power allocation coefficients to be predicted for any given distance set. As a validation data, an average distance set [178 200 222 242 260 279 297 314 330 345 357 372] was used. For the scenario with 2 UEs the first two entry are used, i.e., [178 200], and for the scenario with 12 UEs all the entries of the average distance set were used. Finally, an exhaustive search (optimal), normal equation and the DNN model were executed to obtain power allocation coefficients (α_k's) and related sum capacity is calculated using (5.1). ML and DNN model parameters are given in Table 5.2.

TABLE 5.2: Simulation Parameters

Parameter Name	**Parameter Value**
Number of examples	682
Training examples	618
Testing examples	64
Number of UEs per example	[2 - 12]
Distance	[25 300] (divisible by 25)
Number of layers	3
Accuracy	MAE
Loss	MSE

Fig. 5.3 shows the obtained results. As expected, the sum capacity achieves with exhaustive search for any number of UEs is the highest. The other two methods also give close performance but with significantly lower computational complexity as will be discussed later below. The DNN model has some fluctuations for K is 3 and 8. In contrast, the results of normal equation smoothly resides lower than the exhaustive search with approximately 2 Mbits/sec less sum capacity. Such results foster to trust AI and find encourage to search for more rooms to exploit it.

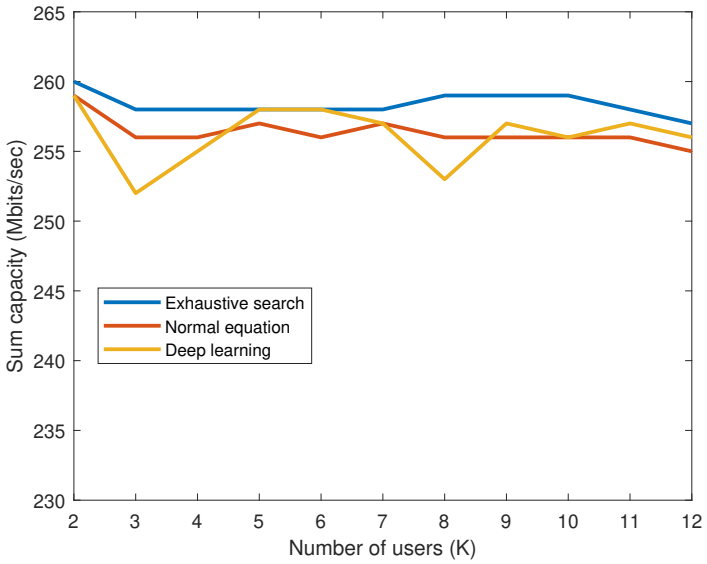

FIGURE 5.3: Sum capacity for different number of UEs with three methods

Time to obtain power allocation coefficients with three methods for different number of UEs are compared in the Table 5.3. It can be seen that the time gradually grew for a large number of UEs. The exhaustive search was the slowest method to obtain power allocation coefficients. It took up to 24.26 ms for the largest set of 12 UEs. In contrast, it took only 0.56 ms for normal equation and only 0.2 ms for DNN to predict power allocation coefficients when $K = 12$.

Figs. 5.4 and 5.5 show the accuracy and loss functions of the DNN method for training and testing data when $K = 2$. The results obtained with the mean absolute error (MAE) and the mean square error (MSE) functions for up to 50 epochs. The batch size was taken

TABLE 5.3: Execution time (ms) of power allocation algorithms.

Number of UEs (K)	Exhaustive search	Normal equation	Deep learning
2	11.89	0.05	0.03
3	12.98	0.08	0.05
4	13.19	0.14	0.07
5	14.55	0.16	0.08
6	15.82	0.21	0.10
7	16.74	0.27	0.12
8	17.93	0.30	0.13
9	18.27	0.39	0.16
10	19.88	0.41	0.16
11	21.33	0.46	0.19
12	24.26	0.56	0.20

as 32 and the DNN rate was set at 0.01. Accuracy results were as high as approximately 0.928 for training and 0.932 for testing by 50 epochs.

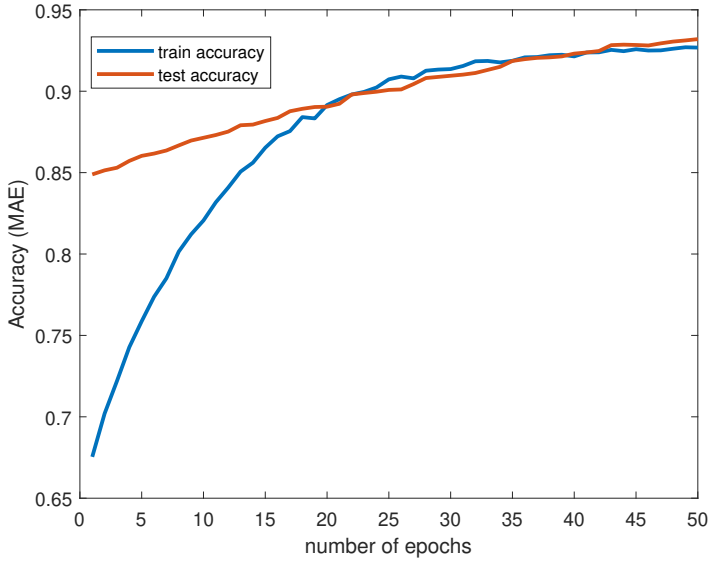

FIGURE 5.4: Accuracy of DNN model

5.7 Chapter Summary

For NOMA networks to achieve a higher capacity than OMA extensively relies on an accurate selection of power allocation coefficients. The existing power allocation strategies (see Section 3.1) are usually computationally heavy or valid under some certain

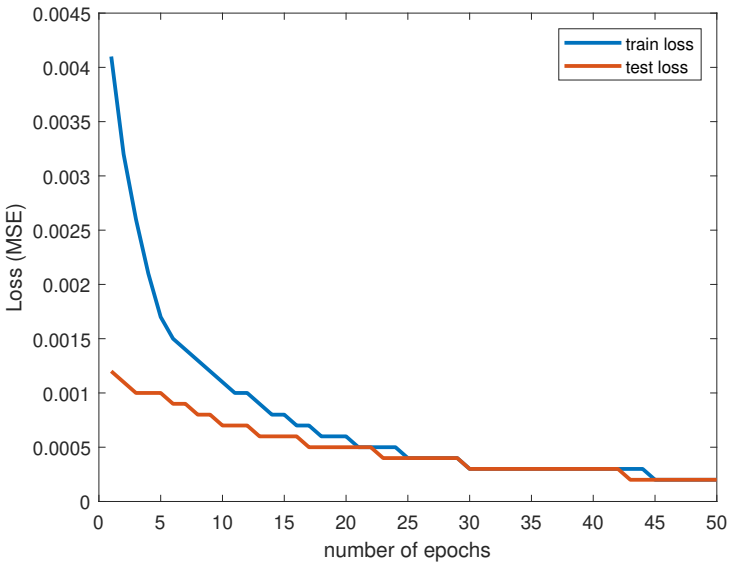

FIGURE 5.5: Loss of DNN model

conditions. This chapter proposed an alternative solution to allocate power in downlink NOMA systems using AI. The proposed methodology was found to be effective in terms of achievable sum capacity and much faster than the existing strategies. The results from this research verified this for the networks with from 2 to 12 UEs. The experiments generated 682 different feature set for each scenario. The measurements used were the mean of the distance vector and then a comparison the results was carried out.

The sum capacity of the scenarios with ML and DL techniques are barely the same and are very close to the optimum power allocation obtained with exhaustive search. Therefore, it can be concluded that the use of AI in power allocation is a reliable tool. Specifically, in a scenario with 50 epochs, the results of MAE function from the training and testing data in DNN showed about 92% of accuracy. Similarly, the loss function of both data sets reached a low of 0.002.

The major achievement of this study was to attain the results with AI in less than a ms. For instance, it took only 0.56 ms for normal equation to obtain the power allocation coefficients for 12 UEs. Also, the DNN coped with the same task for the same set of users in only 0.2 ms.

In real life scenarios, BS processors are overloaded with many tasks to run, e.g., encoding, decoding, modulation/demodulation, precoding, MIMO processing, joint optimization of many physical layer parameters etc. Furthermore, the dynamic nature of mobile cellular networks and high data rate demands lead to an acute update of any computation. In this perspective, NOMA, due to its heavy computational cost, has not yet been exploited in 5G, however it undoubtedly has potential to become the mainframe of upcoming 5G releases and beyond [195]. To ensure this happens, the obstacle of computation cost needs to be resolved and time of any computation has to be reduced as much as possible.

Chapter 6

Conclusion and Future Research

6.1 Conclusion

This chapter concludes the thesis and proposes potential ways to develop the concept beyond. There is a plethora of research, which discuss how NOMA can become a scheme of 5G. Nevertheless, practical issues still prevent the scheme from being adopted. The thesis has included several methods to improve NOMA which assist the scheme in overcoming the challenges it faces in being exploited in 5G.

The thesis has considered a number of architectures, structures for uplink NOMA, structures for downlink NOMA and efficient ways to solve the issues which occur due to computation complexity.

The power domain is a major difference of NOMA from other multiple access schemes. Accurate distribution of power within the signal leads to progress in spectral efficiency. In Section 3.1 an exhaustive search was implemented for optimum power allocation and then the sum capacity for fairness indexes of 0.5, 0.7 and 0.9 was calculated. Then, Section 3.2 discussed a scenario of possible threats to the SIC receiver. SIC is an iterative procedure so the decoding time is expected to be proportional to the number of UEs. We demonstrated that the probability of an attack to a NOMA receiver may be understood from an unusually long decoding time. Finally, in Section 3.3, we compared the processing times to execute decoding procedures of SIC and PIC receivers of NOMA with CPU and GPU. For CPU case we used Java programming language with multi-threading programming

and for the GPU case CUDA parallel programming model was used.

Section 4.1 focused on GPU exploitation at the base stations. NOMA in uplink channel faces extra hard load of signal processing due to the high number of UEs. Firstly it was proposed to break the cell into clusters and keep the number of UEs in a group of 10. Clustering allows to run SIC for a small number of UEs without an unneeded queue to demodulate the message of all UEs in the cell. Parallel computing platform CUDA allows us to run these steps on GPU and achieve 52 times boost in time compared to CPU.

In the second part of the chapter Section 4.2 the power of GPU for NOMA networks with OFDM was demonstrated: the SIC and PIC on CPU and GPU with FFT sizes of 2048 and 4096. We compared SIC and PIC on CPU and on GPU with FFT sizes of 2048 and 4096. The computation time of PIC receiver on GPU was found less than 3 ms for FFT size of 2048 and 7 ms for FFT size of 4096. These figures are 75 and 220 times faster when compared to SIC on CPU. Therefore, practical advice can be recommended as well as working features of GPU in telecommunications. Both of the works shown in Section 4.1 and Section 4.2 consider a large number of users - up to 2000 for the study but with clusters and up to 350 users for the study about NOMA OFDM.

NOMA will most likely address AI to accelerate its performance to be integrated with machine learning and deep learning. After that, it will possibly be included in upcoming releases of 5G. In Section 5.1, the research implemented the exhaustive search of optimum power allocation and then proposed solutions based on linear regression and deep learning. The efficiency of AI has shown to avoid heavy computations and relies on the accurate prediction of the power allocation. The results showed this prediction is possible with the knowledge of users' previous locations and more specifically their distances to the BS. It took less than a 1ms to predict power coefficients with a normal equation of linear regression and about half of that for DNN. Interestingly, exhaustive search spent from approximately 12ms to 24ms to obtain optimum power allocation coefficients for a BS with 2 to 12 UEs. The chapter also compared the sum capacity using power allocation coefficients obtained with all the above methods. Sum capacity with predicted

coefficients was slightly lower than the sum capacity with computed coefficients from exhaustive search (optimal).

6.2 Future Research

The current NOMA related research is limited to algorithms and performance analysis of several scenarios. In this section, the author's interest and scientific curiosity related to the subject in short and long terms is explained.

6.2.1 Short Term

First of all, it is recommended that the single-cell scenario discussed in Section 3.1.1 is revisited with user pairing and the sum capacity results to be compared for large number of users with and without user pairing in both NOMA and OMA systems. Next, it would be noteworthy to enhance the security of SIC Section 3.2 to add the imperfect cancellation scenario that will allow an investigation to be carried out into the residual interference in each iteration. Furthermore, for Section 3.3, it would be beneficial to add a minimum mean square error with parallel interference cancellation (MMSE-PIC) receivers [196] in pairs and then implement it with GPU and multi-threaded CPU.

The study about NOMA clusters in Section 4.1 may be extended with an alternative to SIC receivers like PIC or MMSE-PIC. Then NOMA-OFDM receivers in Section 4.2 may be considered under user pairing and clustered scenario and re-implemented on CPU and GPU hardware.

6.2.2 Long Term

This thesis work has shown that GPUs facilitate the use of computationally complex algorithms in NOMA networks. It would be a step ahead to repeat the GPU experiments on mobile devices with built-in GPUs and evaluate their performances. Moreover, practical implementation in a full-scale cellular system testbed with data transmission and real-time

signal processing in GPU would be a challenging but final work to show the feasibility of using GPUs at NOMA receivers. This can include implementing 5G's synchronization, channel estimation, error correcting processes, control and management signaling mechanisms, de-fragmentation and re-assembly processes at the receiver. Further investigation is required to explore the potential of parallel processing of these tasks and to demonstrate the feasibility of GPUs for 5G receivers. Furthermore, energy consumption of the GPU platforms can be compared with the CPU based systems.

Similar to CUDA, Open ACC may be considered as an alternative which is an open-source platform and is independent of the GPU brand and works well with Intel GPUs. Moreover, multi-threaded CPUs with more than 16 cores may be exploited for the work of the BS receiver for uplink NOMA. In actual cellular networks, movements of UEs within the cell or between the cells require regrouping them in applications with clustering and frequent updates in power allocation. This dynamic behaviour adds to the computational cost. The research on dynamic changes in the cell with experiments on GPU devices leads to practical real-life scenarios.

Interestingly, there are hybrid hardware architectures that use efficiency of DSP, power of GPU and speed of FPGA. The researchers break the computation task into layers and then use suitable device to solve a group of layers with different difficulty. As a result, modified hardware multiplies matrices of large size, trains deep neural networks and decomposes tensor faster than classic CPU or GPU [197–199], . It would be a challenge to modify the structures of the SIC and the PIC receivers and improve the results obtained in this research via such novel hardware architecture.

Bibliography

[1] P. Guturu, "Explosive wireless consumer demand for network bandwidth-fifth generation and beyond [future directions]," *IEEE consumer electronics magazine*, vol. 6, no. 2, pp. 27–31, 2017.

[2] Nokia Solutions and Networks Oy, "Future works 5G use cases and requirements," *Nokia Solutions and Networks Oy*, pp. 9–10, 2014.

[3] Nokia Solutions and Networks Oy, "5G radio access: Requirements, concept and technologies," *Nokia Solutions and Networks Oy*, pp. 2–4, 2014.

[4] Huawei Technologies Co., Ltd., "5G: A technology vision," *Huawei Technologies Co., Ltd.*, pp. 3–5, 2013.

[5] The Mathworks Inc., "5G development with MATLAB," *The Mathworks Inc.*, pp. 28–35, 2017.

[6] A. Checko, H. L. Christiansen, Y. Yan, L. Scolari, G. Kardaras, M. S. Berger, and L. Dittmann, "Cloud ran for mobile networks—a technology overview," *IEEE Communications surveys & tutorials*, vol. 17, no. 1, pp. 405–426, 2014.

[7] A. F. Naguib, A. Paulraj, and T. Kailath, "Capacity improvement of base-station antenna arrays cellular cdma," in *Proceedings of 27th Asilomar Conference on Signals, Systems and Computers*, pp. 1437–1441, IEEE, 1993.

[8] Qualcomm Technologies, Inc., "5G-technology, spectrum, early use case," *Qualcomm Technologies, Inc.*, p. 4, 2018.

[9] P. Popovski *et al.*, "Scenarios, requirements and kpis for 5G mobile and wireless system," *Mobile and wireless communications Enablers for the Twenty-twenty Information Society (METIS)*, p. 19, 2013.

[10] C. Cicconett *et al.*, "5G radio network architecture," *Radio Access and Spectrum FP7 Future Networks Cluster*, p. 2, 01 2014.

[11] J. Jia, T. S. Durrani, and J. Chen, "The innovation waves in mobile telecommunication industry," *IEEE Engineering Management Review*, vol. 46, pp. 63–74, thirdquarter 2018.

[12] L. Dai, B. Wang, Y. Yuan, S. Han, C. I, and Z. Wang, "Non-orthogonal multiple access for 5G: solutions, challenges, opportunities, and future research trends," *IEEE Communications Magazine*, vol. 53, pp. 74–81, Sep. 2015.

[13] A. Benjebbour, Y. Saito, Y. Kishiyama, A. Li, A. Harada, and T. Nakamura, "Concept and practical considerations of non-orthogonal multiple access (NOMA) for future radio access," in *2013 International Symposium on Intelligent Signal Processing and Communication Systems*, pp. 770–774, Nov 2013.

[14] G. Song and X. Wang, "Comparison of interference cancellation schemes for non-orthogonal multiple access system," in *2016 IEEE 83rd Vehicular Technology Conference (VTC Spring)*, pp. 1–5, IEEE, 2016.

[15] Y. Tao, L. Liu, S. Liu, and Z. Zhang, "A survey: Several technologies of non-orthogonal transmission for 5G," *China Communications*, vol. 12, pp. 1–15, Oct 2015.

[16] G. E. Blelloch, "Programming parallel algorithms," *Communications of the ACM*, vol. 39, no. 3, pp. 85–97, 1996.

[17] S. Gochman, A. Mendelson, A. Naveh, and E. Rotem, "Introduction to intel core duo processor architecture.," *Intel Technology Journal*, vol. 10, no. 2, 2006.

[18] Intel, inc., "Intel® Core ™ x-series processors." https://ark.intel.com/products/series/123588/Intel-Core-X-series-Processors. Accessed: 2019-02-05.

[19] M. Kalin, "Concurrent and parallel programming concepts." https://learning.oreilly.com/videos/concurrent-and-parallel/9781771375313, 2015. Accessed: 2019-02-05.

[20] J. de Guzman, "GPU dsp — when you can't have enough cores," 2019. Last accessed 28 April 2020.

[21] F. John, *On performance of GPU and DSP architectures for computationally intensive applications*. PhD thesis, University of Rhode Island, 2013.

[22] Z. Fan, F. Qiu, A. Kaufman, and S. Yoakum-Stover, "GPU cluster for high performance computing," in *Proceedings of the 2004 ACM/IEEE conference on Supercomputing*, p. 47, IEEE Computer Society, 2004.

[23] D. Tse and P. Viswanath, *Fundamentals of wireless communication*. Cambridge university press, 2005.

[24] A. F. Molisch, *Wireless communications*, vol. 34. John Wiley & Sons, 2012.

[25] E. Dahlman, S. Parkvall, and J. Skold, *4G: LTE/LTE-advanced for mobile broadband*. Academic press, 2013.

[26] A. Ghosh, R. Ratasuk, B. Mondal, N. Mangalvedhe, and T. Thomas, "LTE-advanced: next-generation wireless broadband technology," *IEEE wireless communications*, vol. 17, no. 3, pp. 10–22, 2010.

[27] S. Park and D.-H. Cho, "Random linear network coding based on non-orthogonal multiple access in wireless networks," *IEEE Communications Letters*, vol. 19, no. 7, pp. 1273–1276, 2015.

[28] E. Dahlman, S. Parkvall, J. Peisa, and H. Tullberg, "5G evolution and beyond," in *2019 IEEE 20th International Workshop on Signal Processing Advances in Wireless Communications (SPAWC)*, pp. 1–5, IEEE, 2019.

[29] Y. Saito, A. Benjebbour, Y. Kishiyama, and T. Nakamura, "System-level performance evaluation of downlink non-orthogonal multiple access (NOMA)," in *2013 IEEE 24th Annual International Symposium on Personal, Indoor, and Mobile Radio Communications (PIMRC)*, pp. 611–615, IEEE, 2013.

[30] Y. Saito, Y. Kishiyama, A. Benjebbour, T. Nakamura, A. Li, and K. Higuchi, "Non-orthogonal multiple access (NOMA) for cellular future radio access," in *2013 IEEE 77th vehicular technology conference (VTC Spring)*, pp. 1–5, IEEE, 2013.

[31] A. D. Wyner, "Shannon-theoretic approach to a gaussian cellular multiple-access channel," *IEEE Transactions on Information Theory*, vol. 40, no. 6, pp. 1713–1727, 1994.

[32] G. Caire and S. Shamai, "On the achievable throughput of a multiantenna gaussian broadcast channel," *IEEE Transactions on Information Theory*, vol. 49, no. 7, pp. 1691–1706, 2003.

[33] P. Viswanath and D. N. C. Tse, "Sum capacity of the vector gaussian broadcast channel and uplink-downlink duality," *IEEE Transactions on Information Theory*, vol. 49, no. 8, pp. 1912–1921, 2003.

[34] W. Yu and J. M. Cioffi, "Sum capacity of gaussian vector broadcast channels," *IEEE Transactions on Information Theory*, vol. 50, no. 9, pp. 1875–1892, 2004.

[35] J. G. Andrews and T. H. Meng, "Transmit power and other-cell interference reduction via successive interference cancellation with imperfect channel estimation," in *ICC 2001. IEEE International Conference on Communications. Conference Record (Cat. No. 01CH37240)*, vol. 6, pp. 1940–1944, IEEE, 2001.

[36] P. Patel and J. Holtzman, "Analysis of a simple successive interference cancellation scheme in a ds/cdma system," *IEEE journal on selected areas in communications*, vol. 12, no. 5, pp. 796–807, 1994.

[37] J. G. Andrews and T. H. Meng, "Optimum power control for successive interference cancellation with imperfect channel estimation," *IEEE Transactions on Wireless Communications*, vol. 2, no. 2, pp. 375–383, 2003.

[38] N. Docomo, "5G radio access: Requirements, concept and technologies," *white paper, Jul*, 2014.

[39] A. Benjebbovu, A. Li, Y. Saito, Y. Kishiyama, A. Harada, and T. Nakamura, "System-level performance of downlink NOMA for future LTE enhancements," in *2013 IEEE Globecom Workshops (GC Wkshps)*, pp. 66–70, IEEE, 2013.

[40] N. Benvenuto and P. Bisaglia, "Parallel and successive interference cancellation for mc-cdma and their near-far resistance," in *2003 IEEE 58th Vehicular Technology Conference. VTC 2003-Fall (IEEE Cat. No. 03CH37484)*, vol. 2, pp. 1045–1049, IEEE, 2003.

[41] A. Anwar, B.-C. Seet, and X. J. Li, "Pic-based receiver structure for 5G downlink NOMA," in *2015 10th International Conference on Information, Communications and Signal Processing (ICICS)*, pp. 1–5, IEEE, 2015.

[42] Y. Lan, A. Benjebboiu, X. Chen, A. Li, and H. Jiang, "Considerations on downlink non-orthogonal multiple access (NOMA) combined with closed-loop su-mimo," in *2014 8th International Conference on Signal Processing and Communication Systems (ICSPCS)*, pp. 1–5, IEEE, 2014.

[43] T. Manglayev, R. C. Kizilirmak, Y. H. Kho, N. Bazhayev, and I. Lebedev, "NOMA with imperfect sic implementation," in *IEEE EUROCON 2017-17th International Conference on Smart Technologies*, pp. 22–25, IEEE, 2017.

[44] 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)*, pp. 1–7, IEEE, 2017.

[45] Z. Q. Al-Abbasi and D. K. So, "Power allocation for sum rate maximization in non-orthogonal multiple access system," in *2015 IEEE 26th Annual International Symposium on Personal, Indoor, and Mobile Radio Communications (PIMRC)*, pp. 1649–1653, IEEE, 2015.

[46] L. Lei, D. Yuan, C. K. Ho, and S. Sun, "Power and channel allocation for non-orthogonal multiple access in 5G systems: Tractability and computation," *IEEE Transactions on Wireless Communications*, vol. 15, no. 12, pp. 8580–8594, 2016.

[47] S. R. Islam, N. Avazov, O. A. Dobre, and K.-S. 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, 2016.

[48] C. Wang, J. Chen, and Y. Chen, "Power allocation for a downlink non-orthogonal multiple access system," *IEEE Wireless Communications Letters*, vol. 5, pp. 532–535, Oct 2016.

[49] M.-R. Hojeij, J. Farah, C. A. Nour, and C. Douillard, "Resource allocation in downlink non-orthogonal multiple access (NOMA) for future radio access," in *2015 IEEE 81st vehicular technology conference (VTC Spring)*, pp. 1–6, IEEE, 2015.

[50] Y.-N. Lin and D. W. Lin, "On optimal power distribution for successive interference cancellation (sic) for wideband cdma," in *2001 IEEE Third Workshop on Signal Processing Advances in Wireless Communications (SPAWC'01). Workshop Proceedings (Cat. No. 01EX471)*, pp. 38–41, IEEE, 2001.

[51] D. Divsalar, M. K. Simon, and D. Raphaeli, "Improved parallel interference cancellation for cdma," *IEEE Transactions on Communications*, vol. 46, no. 2, pp. 258–268, 1998.

[52] A. Al-Dulaimi, X. Wang, and I. Chih-Lin, *5G Networks: Fundamental Requirements, Enabling Technologies, and Operations Management*. John Wiley & Sons, 2018.

[53] V. K. Garg and T. S. Rappaport, *Wireless network evolution: 2G to 3G*. Prentice Hall PTR, 2001.

[54] J. G. Andrews and T. H. Meng, "Amplitude and phase estimation considerations for asynchronous cdma with successive interference cancellation," in *Vehicular Technology Conference Fall 2000. IEEE VTS Fall VTC2000. 52nd Vehicular Technology Conference (Cat. No. 00CH37152)*, vol. 3, pp. 1211–1215, IEEE, 2000.

[55] 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, 2016.

[56] A. Celik, M.-C. Tsai, R. M. Radaydeh, F. S. Al-Qahtani, and M.-S. Alouini, "Distributed user clustering and resource allocation for imperfect NOMA in heterogeneous networks," *IEEE Transactions on Communications*, vol. 67, no. 10, pp. 7211–7227, 2019.

[57] 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, 2015.

[58] Y. Liu, M. Elkashlan, Z. Ding, and G. K. Karagiannidis, "Fairness of user clustering in mimo non-orthogonal multiple access systems," *IEEE Communications Letters*, vol. 20, no. 7, pp. 1465–1468, 2016.

[59] J.-M. Kang and I.-M. Kim, "Optimal user grouping for downlink NOMA," *IEEE Wireless Communications Letters*, vol. 7, no. 5, pp. 724–727, 2018.

[60] Z. Liu, L. Lei, N. Zhang, G. Kang, and S. Chatzinotas, "Joint beamforming and power optimization with iterative user clustering for miso-NOMA systems," *IEEE Access*, vol. 5, pp. 6872–6884, 2017.

[61] M. A. Sedaghat and R. R. Müller, "On user pairing in uplink NOMA," *IEEE Transactions on Wireless Communications*, vol. 17, no. 5, pp. 3474–3486, 2018.

[62] M. S. Elbamby, M. Bennis, W. Saad, M. Debbah, and M. Latva-Aho, "Resource optimization and power allocation in in-band full duplex-enabled non-orthogonal multiple access networks," *IEEE Journal on Selected Areas in Communications*, vol. 35, no. 12, pp. 2860–2873, 2017.

[63] S. Ali, E. Hossain, and D. I. Kim, "Non-orthogonal multiple access (NOMA) for downlink multiuser mimo systems: User clustering, beamforming, and power allocation," *IEEE access*, vol. 5, pp. 565–577, 2016.

[64] T. Turletti and D. Tennenhouse, "Estimating the computational requirements of a software GSM base station," in *Proceedings of ICC'97-International Conference on Communications*, vol. 1, pp. 169–175, IEEE, 1997.

[65] T. Turletti, H. J. Bentzen, and D. Tennenhouse, "Toward the software realization of a GSM base station," *IEEE Journal on selected areas in communications*, vol. 17, no. 4, pp. 603–612, 1999.

[66] D. Pulley, "Multi-core dsp for base stations: Large and small," in *Proceedings of the 2008 Asia and South Pacific Design Automation Conference*, pp. 389–391, IEEE Computer Society Press, 2008.

[67] B. Peng and J. Chang, "The design and implementation of super base station side l1c system in td-LTE pattern," in *2016 IEEE International Conference on Consumer Electronics-China (ICCE-China)*, pp. 1–5, IEEE, 2016.

[68] J. Huang, Q. Zhang, Q. Li, and J. Qin, "Robust parallel analog function computation via wireless multiple-access mimo channels," *IEEE Signal Processing Letters*, vol. 22, no. 9, pp. 1297–1301, 2015.

[69] M. L. Psiaki, "Real-time generation of bit-wise parallel representations of oversampled prn codes," *IEEE Transactions on Wireless Communications*, vol. 5, no. 3, pp. 487–491, 2006.

[70] T. P. Stefanski, "Fast implementation of fdtd-compatible green's function on multicore processor," *IEEE Antennas and Wireless Propagation Letters*, vol. 11, pp. 81–84, 2012.

[71] A. Somekh-Baruch, "On coding schemes for channels with mismatched decoding," in *2013 IEEE International Symposium on Information Theory*, pp. 96–100, IEEE, 2013.

[72] J. Scarlett, A. Martinez, and A. G. i Fàbregas, "Multiuser random coding techniques for mismatched decoding," *IEEE Transactions on Information Theory*, vol. 62, no. 7, pp. 3950–3970, 2016.

[73] NVIDIA, "*CUDA accelerates applications across a wide range of domains from image processing, to deep learning, numerical analytics and computational science.*," (2019). [Online]. Available: https://www.developer.nvidia.com/cuda-zone.

[74] Q. Zheng, Y. Chen, R. Dreslinski, C. Chakrabarti, A. Anastasopoulos, S. Mahlke, and T. Mudge, "Architecting an LTE base station with graphics processing units," in *SiPS 2013 Proceedings*, pp. 219–224, IEEE, 2013.

[75] K. Li, R. R. Sharan, Y. Chen, T. Goldstein, J. R. Cavallaro, and C. Studer, "Decentralized baseband processing for massive mu-mimo systems," *IEEE Journal on Emerging and Selected Topics in Circuits and Systems*, vol. 7, no. 4, pp. 491–507, 2017.

[76] NVIDIA, "*AI IN TELECOMMUNICATIONS*," (2019). [Online]. Available: https://www.nvidia.com/en-us/industries/telecommunications/.

[77] NVIDIA, "*NETWORK & WIRELESS TELECOM CONFERENCE SESSIONS*," (2019). [Online]. Available: https://www.nvidia.com/en-us/gtc/topics/telecommunication/.

[78] C. Jiang, H. Zhang, Y. Ren, Z. Han, K.-C. Chen, and L. Hanzo, "Machine learning paradigms for next-generation wireless networks," *IEEE Wireless Communications*, vol. 24, no. 2, pp. 98–105, 2016.

[79] T. Gruber, S. Cammerer, J. Hoydis, and S. ten Brink, "On deep learning-based channel decoding," in *2017 51st Annual Conference on Information Sciences and Systems (CISS)*, pp. 1–6, IEEE, 2017.

[80] T. Wang, C.-K. Wen, H. Wang, F. Gao, T. Jiang, and S. Jin, "Deep learning for wireless physical layer: Opportunities and challenges," *China Communications*, vol. 14, no. 11, pp. 92–111, 2017.

[81] S. Shen, C.-J. Chang, and L.-C. Wang, "A cellular neural network and utility-based radio resource scheduler for multimedia cdma communication systems," *IEEE transactions on wireless communications*, vol. 8, no. 11, pp. 5508–5519, 2009.

[82] M. Kim, N.-I. Kim, W. Lee, and D.-H. Cho, "Deep learning-aided scma," *IEEE Communications Letters*, vol. 22, no. 4, pp. 720–723, 2018.

[83] H. Ye, G. Y. Li, and B.-H. Juang, "Power of deep learning for channel estimation and signal detection in ofdm systems," *IEEE Wireless Communications Letters*, vol. 7, no. 1, pp. 114–117, 2017.

[84] Z. M. Fadlullah, F. Tang, B. Mao, N. Kato, O. Akashi, T. Inoue, and K. Mizutani, "State-of-the-art deep learning: Evolving machine intelligence toward tomorrow's

intelligent network traffic control systems," *IEEE Communications Surveys & Tutorials*, vol. 19, no. 4, pp. 2432–2455, 2017.

[85] F. Tang, B. Mao, Z. M. Fadlullah, N. Kato, O. Akashi, T. Inoue, and K. Mizutani, "On removing routing protocol from future wireless networks: A real-time deep learning approach for intelligent traffic control," *IEEE Wireless Communications*, vol. 25, no. 1, pp. 154–160, 2017.

[86] N. Kato, Z. M. Fadlullah, B. Mao, F. Tang, O. Akashi, T. Inoue, and K. Mizutani, "The deep learning vision for heterogeneous network traffic control: Proposal, challenges, and future perspective," *IEEE wireless communications*, vol. 24, no. 3, pp. 146–153, 2016.

[87] B. Mao, Z. M. Fadlullah, F. Tang, N. Kato, O. Akashi, T. Inoue, and K. Mizutani, "Routing or computing? the paradigm shift towards intelligent computer network packet transmission based on deep learning," *IEEE Transactions on Computers*, vol. 66, no. 11, pp. 1946–1960, 2017.

[88] N. Otao, Y. Kishiyama, and K. Higuchi, "Performance of non-orthogonal access with sic in cellular downlink using proportional fair-based resource allocation," in *2012 international symposium on wireless communication systems (ISWCS)*, pp. 476–480, IEEE, 2012.

[89] M.-R. Hojeij, J. Farah, C. A. Nour, and C. Douillard, "New optimal and suboptimal resource allocation techniques for downlink non-orthogonal multiple access," *Wireless Personal Communications*, vol. 87, no. 3, pp. 837–867, 2016.

[90] Y. Sun, D. W. K. Ng, Z. Ding, and R. Schober, "Optimal joint power and subcarrier allocation for mc-NOMA systems," in *2016 IEEE Global Communications Conference (GLOBECOM)*, pp. 1–6, IEEE, 2016.

[91] B. Di, S. Bayat, L. Song, and Y. Li, "Radio resource allocation for downlink non-orthogonal multiple access (NOMA) networks using matching theory," in *2015 IEEE global communications conference (GLOBECOM)*, pp. 1–6, IEEE, 2015.

[92] L. Lei, D. Yuan, C. K. Ho, and S. Sun, "Joint optimization of power and channel allocation with non-orthogonal multiple access for 5G cellular systems," in *2015 IEEE Global Communications Conference (GLOBECOM)*, pp. 1–6, IEEE, 2015.

[93] L. Chen, L. Ma, and Y. Xu, "Proportional fairness-based user pairing and power allocation algorithm for non-orthogonal multiple access system," *IEEE Access*, vol. 7, pp. 19602–19615, 2019.

[94] S. N. Datta and S. Kalyanasundaram, "Optimal power allocation and user selection in non-orthogonal multiple access systems," in *2016 IEEE Wireless Communications and Networking Conference*, pp. 1–6, IEEE, 2016.

[95] 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, 2015.

[96] S. Shi, L. Yang, and H. Zhu, "Outage balancing in downlink nonorthogonal multiple access with statistical channel state information," *IEEE Transactions on Wireless Communications*, vol. 15, no. 7, pp. 4718–4731, 2016.

[97] F. Liu, P. Mähönen, and M. Petrova, "Proportional fairness-based user pairing and power allocation for non-orthogonal multiple access," in *2015 IEEE 26th annual international symposium on personal, indoor, and mobile radio communications (PIMRC)*, pp. 1127–1131, IEEE, 2015.

[98] H. Zhang, D.-K. Zhang, W.-X. Meng, and C. Li, "User pairing algorithm with sic in non-orthogonal multiple access system," in *2016 IEEE International Conference on Communications (ICC)*, pp. 1–6, IEEE, 2016.

[99] M. A. Sedaghat and R. R. Müller, "On user pairing in NOMA uplink," *arXiv preprint arXiv:1707.01846*, 2017.

[100] N. Zhang, J. Wang, G. Kang, and Y. Liu, "Uplink nonorthogonal multiple access in 5G systems," *IEEE Communications Letters*, vol. 20, no. 3, pp. 458–461, 2016.

[101] R. K. Jain, D.-M. W. Chiu, and W. R. Hawe, "A quantitative measure of fairness and discrimination," *Eastern Research Laboratory, Digital Equipment Corporation: Hudson, MA, USA*, pp. 2–7, 1984.

[102] Kcell, "Maintainig a leading position," *annual report and accounts*, 2018.

[103] P. Schneider and G. Horn, "Towards 5G security," in *2015 IEEE Trustcom/BigDataSE/ISPA*, vol. 1, pp. 1165–1170, IEEE, 2015.

[104] E. Dahlman, G. Mildh, S. Parkvall, J. Peisa, J. Sachs, and Y. Selén, "5G radio access," *Ericsson review*, vol. 6, no. 1, 2014.

[105] N. Alliance, "5G white paper," *Next generation mobile networks, white paper*, vol. 1, 2015.

[106] A. J. Viterbi, "Very low rate convolutional codes for maximum theoretical performance of spread-spectrum multiple-access channels," in *The Foundations Of The Digital Wireless World: Selected Works of AJ Viterbi*, pp. 115–123, World Scientific, 2010.

[107] G. B. Satrya and S. Y. Shin, "Security enhancement to successive interference cancellation algorithm for non-orthogonal multiple access (NOMA)," in *2017 IEEE 28th Annual International Symposium on Personal, Indoor, and Mobile Radio Communications (PIMRC)*, pp. 1–5, IEEE, 2017.

[108] Z. Ding, X. Lei, G. K. Karagiannidis, R. Schober, J. Yuan, and V. K. Bhargava, "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, 2017.

[109] N. Zhao, F. R. Yu, M. Li, Q. Yan, and V. C. Leung, "Physical layer security issues in interference-alignment-based wireless networks," *IEEE Communications Magazine*, vol. 54, no. 8, pp. 162–168, 2016.

[110] B. He, A. Liu, N. Yang, and V. K. Lau, "On the design of secure non-orthogonal multiple access systems," *IEEE Journal on Selected Areas in Communications*, vol. 35, no. 10, pp. 2196–2206, 2017.

[111] H. Lei, J. Zhang, K.-H. Park, P. Xu, I. S. Ansari, G. Pan, B. Alomair, and M.-S. Alouini, "On secure NOMA systems with transmit antenna selection schemes," *IEEE Access*, vol. 5, pp. 17450–17464, 2017.

[112] Y. Zhang, H.-M. Wang, Q. Yang, and Z. Ding, "Secrecy sum rate maximization in non-orthogonal multiple access," *IEEE Communications Letters*, vol. 20, no. 5, pp. 930–933, 2016.

[113] Y. Cao, N. Zhao, Y. Chen, M. Jin, L. Fan, Z. Ding, and F. R. Yu, "Privacy protection via beamforming optimization in miso NOMA networks," in *2018 10th International Conference on Wireless Communications and Signal Processing (WCSP)*, pp. 1–6, IEEE, 2018.

[114] I. Lebedev, I. Krivtsova, V. Korzhuk, N. Bazhayev, M. Sukhoparov, S. Pecherkin, and K. Salakhutdinova, "The analysis of abnormal behavior of the system local segment on the basis of statistical data obtained from the network infrastructure monitoring," in *Internet of Things, Smart Spaces, and Next Generation Networks and Systems*, pp. 503–511, Springer, 2016.

[115] N. Bazhayev, I. Lebedev, I. Krivtsova, and I. Zikratov, "Research availability of devices based on wireless networks," in *2016 18th Conference of Open Innovations Association and Seminar on Information Security and Protection of Information Technology (FRUCT-ISPIT)*, pp. 27–32, IEEE, 2016.

[116] N. Bazhayev, I. Lebedev, I. Krivtsova, M. Sukhoparov, K. Salakhutdinova, A. Davydov, and Y. Shaparenko, "Evaluation of the available wireless remote devices subject to the information impact," in *2016 IEEE 10th International Conference on Application of Information and Communication Technologies (AICT)*, pp. 1–6, IEEE, 2016.

[117] N. Bazhayev, I. Lebedev, V. Korzhuk, and I. Zikratov, "Monitoring of the information security of wireless remote devices," in *2015 9th International Conference on Application of Information and Communication Technologies (AICT)*, pp. 233–236, IEEE, 2015.

[118] T. Manglayev, R. C. Kizilirmak, and Y. H. Kho, "Optimum power allocation for non-orthogonal multiple access (NOMA)," in *2016 IEEE 10th International Conference on Application of Information and Communication Technologies (AICT)*, pp. 1–4, IEEE, 2016.

[119] 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)*, pp. 781–785, Aug 2014.

[120] B. Xia, J. Wang, K. Xiao, Y. Gao, Y. Yao, and S. Ma, "Outage performance analysis for the advanced sic receiver in wireless NOMA systems," *IEEE Transactions on Vehicular Technology*, vol. 67, no. 7, pp. 6711–6715, 2018.

[121] Y. Gao, B. Xia, K. Xiao, Z. Chen, X. Li, and S. Zhang, "Theoretical analysis of the dynamic decode ordering sic receiver for uplink NOMA systems," *IEEE Communications Letters*, vol. 21, no. 10, pp. 2246–2249, 2017.

[122] M. Chen and A. Burr, "Multiuser detection for uplink non-orthogonal multiple access system," *IET Communications*, vol. 13, no. 19, pp. 3222–3228, 2019.

[123] V. Ghazi-Moghadam, L. B. Nelson, and M. Kaveh, "Parallel interference cancellation for cdma systems," in *PROCEEDINGS OF THE ANNUAL ALLERTON*

CONFERENCE ON COMMUNICATION CONTROL AND COMPUTING, vol. 33, pp. 216–224, Citeseer, 1995.

[124] D. Guo, L. K. Rasmussen, S. Sun, and T. J. Lim, "A matrix-algebraic approach to linear parallel interference cancellation in cdma," *IEEE Transactions on Communications*, vol. 48, no. 1, pp. 152–161, 2000.

[125] H. Yan and S. Roy, "Parallel interference cancellation for uplink multirate overlay cdma channels," *IEEE transactions on communications*, vol. 53, no. 1, pp. 152–161, 2005.

[126] Oracle, inc., "Class thread." https://docs.oracle.com/javase/8/docs/api/index.html?java/lang/Thread.html. Accessed: 2019-08-19.

[127] B. Kimy, S. Lim, H. Kim, S. Suh, J. Kwun, S. Choi, C. Lee, S. Lee, and D. Hong, "Non-orthogonal multiple access in a downlink multiuser beamforming system," in *MILCOM 2013-2013 IEEE Military Communications Conference*, pp. 1278–1283, IEEE, 2013.

[128] F. Alavi, K. Cumanan, Z. Ding, and A. G. Burr, "Beamforming techniques for nonorthogonal multiple access in 5G cellular networks," *IEEE Transactions on Vehicular Technology*, vol. 67, no. 10, pp. 9474–9487, 2018.

[129] K. Higuchi and A. Benjebbour, "Non-orthogonal multiple access (NOMA) with successive interference cancellation for future radio access," *IEICE Transactions on Communications*, vol. 98, no. 3, pp. 403–414, 2015.

[130] A. Li, A. Benjebbour, X. Chen, H. Jiang, and H. Kayama, "Uplink non-orthogonal multiple access (NOMA) with single-carrier frequency division multiple access (sc-fdma) for 5G systems," *IEICE Transactions on Communications*, vol. 98, no. 8, pp. 1426–1435, 2015.

[131] M. M. Al-Wani, A. Sali, N. K. Noordin, S. J. Hashim, C. Y. Leow, and I. Krikidis, "Robust beamforming and user clustering for guaranteed fairness in downlink NOMA with partial feedback," *IEEE Access*, vol. 7, pp. 121599–121611, 2019.

[132] K. Higuchi and Y. Kishiyama, "Non-orthogonal access with random beamforming and intra-beam sic for cellular mimo downlink," in *2013 IEEE 78th Vehicular Technology Conference (VTC Fall)*, pp. 1–5, IEEE, 2013.

[133] Z. Ding, F. Adachi, and H. V. Poor, "The application of mimo to non-orthogonal multiple access," *IEEE Transactions on Wireless Communications*, vol. 15, no. 1, pp. 537–552, 2015.

[134] R. Vannithamby and S. Talwar, *Towards 5G: Applications, requirements and candidate technologies*. John Wiley & Sons, 2017.

[135] S. Ali, E. Hossain, and D. I. Kim, "Non-orthogonal multiple access (NOMA) for downlink multiuser mimo systems: User clustering, beamforming, and power allocation," *IEEE access*, vol. 5, pp. 565–577, 2017.

[136] H. Tabassum, E. Hossain, and J. Hossain, "Modeling and analysis of uplink non-orthogonal multiple access in large-scale cellular networks using poisson cluster processes," *IEEE Transactions on Communications*, vol. 65, no. 8, pp. 3555–3570, 2017.

[137] N. Wilt, *The cuda handbook: A comprehensive guide to GPU programming*. Pearson Education, 2013.

[138] N. I. Miridakis and D. D. Vergados, "A survey on the successive interference cancellation performance for single-antenna and multiple-antenna ofdm systems," *IEEE Communications Surveys & Tutorials*, vol. 15, no. 1, pp. 312–335, 2012.

[139] H. S. Ghazi and K. W. Wesołowski, "Improved detection in successive interference cancellation NOMA ofdm receiver," *IEEE Access*, vol. 7, pp. 103325–103335, 2019.

[140] W. Xu, X. Zhou, C.-H. Lee, Z. Feng, and J. Lin, "Energy-efficient joint sensing duration, detection threshold, and power allocation optimization in cognitive ofdm systems," *IEEE transactions on wireless communications*, vol. 15, no. 12, pp. 8339–8352, 2016.

[141] B. Di, L. Song, and Y. Li, "Sub-channel assignment, power allocation, and user scheduling for non-orthogonal multiple access networks," *IEEE Transactions on Wireless Communications*, vol. 15, no. 11, pp. 7686–7698, 2016.

[142] Y. Liu, Z. Ding, M. Elkashlan, and J. Yuan, "Nonorthogonal multiple access in large-scale underlay cognitive radio networks," *IEEE Transactions on Vehicular Technology*, vol. 65, no. 12, pp. 10152–10157, 2016.

[143] J. Dai and S. Wang, "Clustering-based spectrum sharing strategy for cognitive radio networks," *IEEE Journal on Selected Areas in Communications*, vol. 35, no. 1, pp. 228–237, 2016.

[144] W. Xu, X. Li, C.-H. Lee, M. Pan, and Z. Feng, "Joint sensing duration adaptation, user matching, and power allocation for cognitive ofdm-NOMA systems," *IEEE Transactions on Wireless Communications*, vol. 17, no. 2, pp. 1269–1282, 2017.

[145] M. Agiwal, A. Roy, and N. Saxena, "Next generation 5G wireless networks: A comprehensive survey," *IEEE Communications Surveys & Tutorials*, vol. 18, no. 3, pp. 1617–1655, 2016.

[146] A. Li, Y. Lan, X. Chen, and H. Jiang, "Non-orthogonal multiple access (NOMA) for future downlink radio access of 5G," *China Communications*, vol. 12, no. Supplement, pp. 28–37, 2015.

[147] Q. C. Li, H. Niu, A. T. Papathanassiou, and G. Wu, "5G network capacity: Key elements and technologies," *IEEE Vehicular Technology Magazine*, vol. 9, no. 1, pp. 71–78, 2014.

[148] M. K. Varanasi and B. Aazhang, "Multistage detection in asynchronous code-division multiple-access communications," *IEEE Transactions on communications*, vol. 38, no. 4, pp. 509–519, 1990.

[149] S. C. Kim and S. S. Bhattacharyya, "A wideband front-end receiver implementation on GPUs," *IEEE Transactions on Signal Processing*, vol. 64, no. 10, pp. 2602–2612, 2016.

[150] R. C. Kizilirmak, *Towards 5G Wireless Networks: A Physical Layer Perspective*. BoD–Books on Demand, 2016.

[151] A. Goldsmith, *Wireless communications*. Cambridge university press, 2005.

[152] R. M. Buehrer, N. S. Correal-Mendoza, and B. D. Woerner, "A simulation comparison of multiuser receivers for cellular cdma," *IEEE Transactions on Vehicular Technology*, vol. 49, no. 4, pp. 1065–1085, 2000.

[153] Y. L. Guan, *Recent Advances in Information, Communications and Signal Processing*. River Publishers, 2018.

[154] A. Kaul and B. D. Woerner, "Analytic limits on performance of adaptive multistage interference cancellation for cdma," *Electronics Letters*, vol. 30, no. 25, pp. 2093–2095, 1994.

[155] W. Shin, M. Vaezi, B. Lee, D. J. Love, J. Lee, and H. V. Poor, "Non-orthogonal multiple access in multi-cell networks: Theory, performance, and practical challenges," *IEEE Communications Magazine*, vol. 55, no. 10, pp. 176–183, 2017.

[156] M. Sundararajan and U. Govindaswamy, "Multicarrier spread spectrum modulation schemes and efficient fft algorithms for cognitive radio systems," *Electronics*, vol. 3, no. 3, pp. 419–443, 2014.

[157] NVIDIA Corporation, "*CUDA Toolkit 4.2 CUFFT Library. Programming Guide.*," (2012). [Online]. Available: https://developer.download.nvidia.

com/compute/DevZone/docs/html/CUDALibraries/doc/CUFFT_Library.pdf.

[158] U. A. Acar and G. Blelloch, "Algorithm design: Parallel and sequential," 2016.

[159] W. Roh, J.-Y. Seol, J. Park, B. Lee, J. Lee, Y. Kim, J. Cho, K. Cheun, and F. Aryanfar, "Millimeter-wave beamforming as an enabling technology for 5G cellular communications: Theoretical feasibility and prototype results," *IEEE communications magazine*, vol. 52, no. 2, pp. 106–113, 2014.

[160] K. Yu, T. Sato, *et al.*, "Modeling and analysis of error process in 5G wireless communication using two-state markov chain," *IEEE Access*, vol. 7, pp. 26391–26401, 2019.

[161] T. S. Rappaport, "5G millimeter wave wireless: Trials, testimonies, and target rollouts," in *IEEE Infocom*, 2018.

[162] Y. Li and G. A. A. Baduge, "Underlay spectrum-sharing massive mimo NOMA," *IEEE Communications Letters*, vol. 23, no. 1, pp. 116–119, 2018.

[163] T. Manglayev, R. C. Kizilirmak, Y. H. Kho, and N. A. W. A. Hamid, "GPU accelerated successive interference cancellation for NOMA uplink with user clustering," *Wireless Personal Communications*, vol. 103, no. 3, pp. 2391–2400, 2018.

[164] A. Carroll, G. Heiser, *et al.*, "An analysis of power consumption in a smartphone.," in *USENIX annual technical conference*, vol. 14, pp. 21–21, Boston, MA, 2010.

[165] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[166] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, pp. 1097–1105, 2012.

[167] A. H. Marblestone, G. Wayne, and K. P. Kording, "Toward an integration of deep learning and neuroscience," *Frontiers in computational neuroscience*, vol. 10, p. 94, 2016.

[168] G. Gui, H. Huang, Y. Song, and H. Sari, "Deep learning for an effective nonorthogonal multiple access scheme," *IEEE Transactions on Vehicular Technology*, vol. 67, no. 9, pp. 8440–8450, 2018.

[169] H. Ye, G. Y. Li, and B.-H. F. Juang, "Deep reinforcement learning based resource allocation for v2v communications," *IEEE Transactions on Vehicular Technology*, vol. 68, no. 4, pp. 3163–3173, 2019.

[170] L. Xiao, Y. Li, C. Dai, H. Dai, and H. V. Poor, "Reinforcement learning-based NOMA power allocation in the presence of smart jamming," *IEEE Transactions on Vehicular Technology*, vol. 67, no. 4, pp. 3377–3389, 2017.

[171] Y. Wei, F. R. Yu, M. Song, and Z. Han, "User scheduling and resource allocation in hetnets with hybrid energy supply: An actor-critic reinforcement learning approach," *IEEE Transactions on Wireless Communications*, vol. 17, no. 1, pp. 680–692, 2017.

[172] A. Chiumento, C. Desset, S. Pollin, L. Van der Perre, and R. Lauwereins, "Impact of csi feedback strategies on LTE downlink and reinforcement learning solutions for optimal allocation," *IEEE Transactions on Vehicular Technology*, vol. 66, no. 1, pp. 550–562, 2016.

[173] J. Cui, Z. Ding, and P. Fan, "The application of machine learning in mmwave-NOMA systems," in *2018 IEEE 87th Vehicular Technology Conference (VTC Spring)*, pp. 1–6, IEEE, 2018.

[174] J. Cui, Z. Ding, P. Fan, and N. Al-Dhahir, "Unsupervised machine learning-based user clustering in millimeter-wave-NOMA systems," *IEEE Transactions on Wireless Communications*, vol. 17, no. 11, pp. 7425–7440, 2018.

[175] J. Luo, J. Tang, D. K. So, G. Chen, K. Cumanan, and J. A. Chambers, "A deep learning-based approach to power minimization in multi-carrier NOMA with swipt," *IEEE Access*, vol. 7, pp. 17450–17460, 2019.

[176] Y. Bengio, A. Courville, and P. Vincent, "Representation learning: A review and new perspectives," *IEEE transactions on pattern analysis and machine intelligence*, vol. 35, no. 8, pp. 1798–1828, 2013.

[177] Y.-F. Liu and Y.-H. Dai, "On the complexity of joint subcarrier and power allocation for multi-user ofdma systems," *IEEE transactions on Signal Processing*, vol. 62, no. 3, pp. 583–596, 2013.

[178] M. F. Hanif, Z. Ding, T. Ratnarajah, and G. K. Karagiannidis, "A minorization-maximization method for optimizing sum rate in the downlink of non-orthogonal multiple access systems," *IEEE Transactions on Signal Processing*, vol. 64, no. 1, pp. 76–88, 2015.

[179] J. Zhu, J. Wang, Y. Huang, S. He, X. You, and L. Yang, "On optimal power allocation for downlink non-orthogonal multiple access systems," *IEEE Journal on Selected Areas in Communications*, vol. 35, no. 12, pp. 2744–2757, 2017.

[180] J. Choi, "Power allocation for max-sum rate and max-min rate proportional fairness in NOMA," *IEEE Communications Letters*, vol. 20, no. 10, pp. 2055–2058, 2016.

[181] J. Cui, Z. Ding, and P. Fan, "A novel power allocation scheme under outage constraints in NOMA systems," *IEEE Signal Processing Letters*, vol. 23, no. 9, pp. 1226–1230, 2016.

[182] Y. Sun, D. W. K. Ng, Z. Ding, and R. Schober, "Optimal joint power and subcarrier allocation for full-duplex multicarrier non-orthogonal multiple access systems," *IEEE Transactions on Communications*, vol. 65, no. 3, pp. 1077–1091, 2017.

[183] Y. Hayashi, Y. Kishiyama, and K. Higuchi, "Investigations on power allocation among beams in non-orthogonal access with random beamforming and intra-beam

sic for cellular mimo downlink," in *2013 IEEE 78th Vehicular Technology Conference (VTC Fall)*, pp. 1–5, IEEE, 2013.

[184] D. Zhai, R. Zhang, L. Cai, B. Li, and Y. Jiang, "Energy-efficient user scheduling and power allocation for NOMA-based wireless networks with massive iot devices," *IEEE Internet of Things Journal*, vol. 5, no. 3, pp. 1857–1868, 2018.

[185] M. Liu, T. Song, and G. Gui, "Deep cognitive perspective: Resource allocation for NOMA-based heterogeneous iot with imperfect sic," *IEEE Internet of Things Journal*, vol. 6, no. 2, pp. 2885–2894, 2018.

[186] R. Amiri, H. Mehrpouyan, L. Fridman, R. K. Mallik, A. Nallanathan, and D. Matolak, "A machine learning approach for power allocation in hetnets considering qos," in *2018 IEEE International Conference on Communications (ICC)*, pp. 1–7, IEEE, 2018.

[187] M. Liu, T. Song, L. Zhang, and G. Gui, "Resource allocation for NOMA based heterogeneous iot with imperfect sic: A deep learning method," in *2018 IEEE 29th Annual International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC)*, pp. 1440–1446, IEEE, 2018.

[188] F. Jameel, W. U. Khan, Z. Chang, T. Ristaniemi, and J. Liu, "Secrecy analysis and learning-based optimization of cooperative NOMA swipt systems," in *2019 IEEE International Conference on Communications Workshops (ICC Workshops)*, pp. 1–6, IEEE, 2019.

[189] Y. Sun, Y. Wang, J. Jiao, S. Wu, and Q. Zhang, "Deep learning-based long-term power allocation scheme for NOMA downlink system in s-iot," *IEEE Access*, vol. 7, pp. 86288–86296, 2019.

[190] Y. Zhang, X. Wang, and Y. Xu, "Energy-efficient resource allocation in uplink NOMA systems with deep reinforcement learning," in *2019 11th International*

Conference on Wireless Communications and Signal Processing (WCSP), pp. 1–6, IEEE, 2019.

[191] F. Hussain, S. A. Hassan, R. Hussain, and E. Hossain, "Machine learning for resource management in cellular and iot networks: Potentials, current solutions, and open challenges," *arXiv preprint arXiv:1907.08965*, 2019.

[192] W. Kim, Y. Ahn, and B. Shim, "Deep neural network based active user detection for grant-free NOMA systems," *arXiv preprint arXiv:1912.11782*, 2019.

[193] W. Kahan, "Ieee standard 754 for binary floating-point arithmetic," *Lecture Notes on the Status of IEEE*, vol. 754, no. 94720-1776, p. 11, 1996.

[194] Y. Xin, Y.-H. Nam, Y. Li, and J. C. Zhang, "Reduced complexity precoding and scheduling algorithms for full-dimension mimo systems," in *2014 IEEE Global Communications Conference*, pp. 4858–4863, IEEE, 2014.

[195] B. Makki, K. Chitti, A. Behravan, and M.-S. Alouini, "A survey of NOMA: Current status and open research challenges," *arXiv preprint arXiv:1912.10561 [cs.IT]*, 2019.

[196] Z. Wu, K. Lu, C. Jiang, and X. Shao, "Comprehensive study and comparison on 5G NOMA schemes," *IEEE Access*, vol. 6, pp. 18511–18519, 2018.

[197] R. Acosta-Quiñonez and R. Rodriguez-Avila, "Tensor decomposition over a fast-prototyping hcp composed by cpu-GPU," in *2019 IEEE Latin-American Conference on Communications (LATINCOM)*, pp. 1–6, IEEE, 2019.

[198] T. Suzuki, S.-Y. Kim, J.-I. Kani, and J. Terada, "Real-time implementation of coherent receiver dsp adopting stream split assignment on GPU for flexible optical access systems," *Journal of Lightwave Technology*, 2019.

[199] Y. Kim and Y. Park, "Blockwise weighted least square active noise control for cpu-GPU architecture," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 951–963, 2020.