OPTIMIZATION OF HANDOVER, SURVIVABILITY, MULTI-CONNECTIVITY AND SECURE SLICING IN 5G CELLULAR NETWORKS USING MATRIX EXPONENTIAL MODELS AND MACHINE LEARNING

A Dissertation
IN
Telecommunications and Computer Networking
and
Electrical and Computer Engineering

Presented to the Faculty of the University
of Missouri–Kansas City in partial fulfillment of
the requirements for the degree

DOCTOR OF PHILOSOPHY

by
RAHUL ARUN PAROPKARI

PhD, University of Missouri-Kansas City

Kansas City, Missouri
2022

OPTIMIZATION OF HANDOVER, SURVIVABILITY, MULTI-CONNECTIVITY AND SECURE SLICING IN 5G CELLULAR NETWORKS USING MATRIX EXPONENTIAL MODELS AND MACHINE LEARNING

Rahul Arun Paropkari, Candidate for the Doctor of Philosophy Degree

University of Missouri–Kansas City, 2022

ABSTRACT

This works proposes optimization of cellular handovers, cellular network survivability modeling, multi-connectivity and secure network slicing using matrix exponentials and machine learning techniques. We propose matrix exponential (ME) modeling of handover arrivals with the potential to much more accurately characterize arrivals and prioritize resource allocation for handovers, especially handovers for emergency or public safety needs. With the use of a 'B' matrix for representing a handover arrival, we have a rich set of dimensions to model system handover behavior. We can study multiple parameters and the interactions between system events along with the user mobility, which would trigger a handoff in any given scenario. Additionally, unlike any traditional handover improvement scheme, we develop a 'Deep-Mobility' model by implementing a deep learning neural network (DLNN) to manage network mobility, utilizing in-network

deep learning and prediction. We use the radio and the network key performance indicators (KPIs) to train our model to analyze network traffic and handover requirements.

Cellular network design must incorporate disaster response, recovery and repair scenarios. Requirements for high reliability and low latency often fail to incorporate network survivability for mission critical and emergency services. Our Matrix Exponential (ME) model shows how survivable networks can be designed based on controlling numbers of crews, times taken for individual repair stages, and the balance between fast and slow repairs. Transient and the steady state representations of system repair models, namely, fast and slow repairs for networks consisting of multiple repair crews have been analyzed. Failures are exponentially modeled as per common practice, but ME distributions describe the more complex recovery processes.

In some mission critical communications, the availability requirements may exceed five or even six nines (99.9999%). To meet such a critical requirement and minimize the impact of mobility during handover, a Fade Duration Outage Probability (FDOP) based multiple radio link connectivity handover method has been proposed. By applying such a method, a high degree of availability can be achieved by utilizing two or more uncorrelated links based on minimum FDOP values. Packet duplication (PD) via multi-connectivity is a method of compensating for lost packets on a wireless channel. Utilizing two or more uncorrelated links, a high degree of availability can be attained with this strategy. However, complete packet duplication is inefficient and frequently unnecessary. We

provide a novel adaptive fractional packet duplication (A-FPD) mechanism for enabling and disabling packet duplication based on a variety of parameters.

We have developed a 'DeepSlice' model by implementing Deep Learning (DL) Neural Network to manage network load efficiency and network availability, utilizing in-network deep learning and prediction. Our Neural Network based 'Secure5G' Network Slicing model will proactively detect and eliminate threats based on incoming connections before they infest the 5G core network elements. These will enable the network operators to sell network slicing as-a-service to serve diverse services efficiently over a single infrastructure with higher level of security and reliability.

APPROVAL PAGE

The faculty listed below, appointed by the Dean of the School of Graduate Studies, have examined a dissertation titled "Optimization of Handover, Survivability, Multi-Connectivity and Secure Slicing in 5G Cellular Networks using Matrix Exponential Models and Machine Learning," presented by Rahul Arun Paropkari, candidate for the Doctor of Philosophy degree, and certify that in their opinion it is worthy of acceptance.

Supervisory Committee

Cory Beard, Ph.D., Committee Chair
Department of Computer Science & Electrical Engineering

Ghulam Chaudhry, Ph.D.
Department of Computer Science & Electrical Engineering

Deep Medhi, Ph.D.
Department of Computer Science & Electrical Engineering

Masud Chowdhury, Ph.D.
Department of Computer Science & Electrical Engineering

Appie Van de Liefvoort, Ph.D.
Department of Computer Science & Electrical Engineering

CONTENTS

ILLUSTRATIONS

TABLES

ACKNOWLEDGEMENTS

I would like to thank my academic and research advisor Dr. Cory Beard for his dedicated support and guidance throughout my PhD process. He continuously provided encouragement and was always willing and enthusiastic to assist in any way. He has been my constant pillar of support for the last 13 years in my academics and personal endeavors. He's been extremely patient with me and ensured that I always had all the resources available for completing my research work successfully.

I would like to thank Dr. Appie Van De Liefvoort for all his support in publishing some of my research work. I have gained a lot of technical knowledge during his guidance in this work. I would also like to thank all of my committee members for being available and providing me with the right opportunities to excel in this degree program. My PhD and two masters degree from UMKC would not have been possible without the encouragement and support of all the faculty members, academic advisors, and my peers.

What I am today is only because of my parents, my father, the Late Dr. Arun Paropkari and my mother Dr. Anjalee Paropkari who have always encouraged me to pursue higher education with complete emotional and financial support. My brother, Mithil Paropkari, for his constant support and agreement with everything I do. I also want to thank my research partners and extended family members Anurag Thantharate, Vijay Walunj, Rohit Abhishek, Poonam Kankariya and Priyanka Walunj for always being there during good and tough times.

CHAPTER 1

INTRODUCTION

The mobile industry has exponentially grown in the past few years. The number of mobile devices has been growing quickly and will continue to do so. Mobile device communication has become an indispensable part of daily life. Cellular Handovers or the Cellular Mobility is very important for the seamless connectivity and there are multiple factors that will be impacting the handover decision. The wireless communications ecosystem has evolved over time and there has been an exponential growth in the number of base stations (BSs) being installed for increased coverage and enhanced network usage. The mobile connectivity index in USA alone grew from 74.1 in 2014 to 85 in 2019 [1] and has had continuous growth since then. Network Availability is mandatory requirement for the future wireless networks as more critical applications will be relying on the coverage of the cellular networks.

In the next generation of wireless networks, that includes dense cellular deployments, handovers will continue to play an important role. Any user will have to be guaranteed a continuous connection, and at the same time, too many handoffs will also have to be avoided. The 5G proposes presence of a heterogeneous environment where a user will seamlessly have to undergo a vertical handover without any service interruption times. These vertical handoffs need to be carried out seamlessly and precisely when required. Dual or multiple connectivity helps meet the mobility requirements for certain essential 5G use cases and ensures the user's connection to one or more radio links. An important

performance indicator for wireless networks is the signal-to-interference-plus-noise ratio (SINR). Packet duplication (PD) via multi-connectivity is a method of compensating for lost packets on a wireless channel.

The research is focused on four separate streams and each one is discussed in detail in their appropriate chapters. The second chapter is about modeling of handovers using two methods; Matrix Exponential and Deep Learning Neural Networks. The third chapter discusses about Survivability modeling of a Cellular Base Station (BS) network after a catastrophic disaster which renders all the BS in an inactive state. A Matrix exponential model has been implemented to model the disaster recovery process of a network of BS. Fourth chapter is about the Fade Duration Outage Probability based HO scheme and Fractional Packet Duplication (FPD). The PD using Multi-Connectivity is further explored using multiple schemes involved in Adaptive-FPD (A-FPD). Chapter 5 shows our proposed DeepSlice and Secure5G models.

1.1 5G Cellular Mobility

In some industries like the public safety, emergency, medical, or energy sectors, mobile devices provide great capabilities but are required to be highly reliable at the same time. In order to provide uninterrupted services to these devices on the wireless network, the infrastructure has struggled over time and has been just about up to the mark. 4G-LTE was able to cater to most needs supporting 720p on video but most content available going forward will be Full HD, UHD offering 4K and 8K video quality. Augmented Reality (AR), Virtual Reality (VR) also add to the bandwidth demands on some online shopping

applications to provide a rich shopping experience. In Chapter 2, we show the fundamental modeling approach and demonstrate usefulness of the model while investigating impacts and sensitivities of certain key parameters and KPIs from the user equipment (UE) and network side.

1.2 Network Survivability

Industries like medicine, banking, energy, and education are relying on wireless devices with demands for high speed connectivity at all times. Increase in network usage has also led to a higher use of time-sensitive applications requiring a very high level of reliability and low latency as defined as per the 5G standards [2]. Our model in Chapter 3 is evaluated for varying number of repair crews, base stations, repair models and squared coefficient of variation values. This model is scalable for larger networks, calculates the restoration and network availability times, consists of asymptotic approximations to estimate network availability and determines the optimal number repair crews required.

1.3 Multi-Connectivity and Fractional Packet Duplication

Service interruption is not acceptable for certain applications, and handovers are the means to avoid loss in connectivity. Fade Duration Outage Probability defines a time over which a communication will fail if a fade persists too long. For example, dropping successive packets can cause lost source relationships and channel coding fails after too long of sequences of errors. FDOP provides a direct relationship with quality of a connection. We show that FDOP may result in a reduced range of coverage, but also more time for the handover process. FDOP is also very helpful for multi-connectivity cases. We

introduce a novel fractional packet duplication process along with FDOP to only duplicate enough packets over multiple connections to meet outage requirements. We evaluate and contrast several packet duplication scenarios and present our simulation results for multi-connectivity. Our technique merely duplicates enough packets across multiple connections to meet the outage criteria without compromising on any limited radio resources. Full packet duplication over multiple links is wasteful and frequently unnecessary so we demonstrate our Adaptive Fractional Packet Duplication (A-FPD) in Chapter 4.

1.4 DeepSlice and Secure5G Network Slicing

Many emerging technologies have taken ablaze the telecom industry by enabling new business models and providing customers a different experience. Networks have evolved with the introduction of programmable systems like the Software Defined Networks (SDN) and Network Function Virtualization (NFV) and have benefited ever since their implementation. Some critical services that 5G networks would encapsulate are autonomous driving, enterprise business models, AR-VR solutions, industrial automation, remote monitoring, smart health, smart cities, and many more. The Third Generation Partnership Project (3GPP) considers network slicing a key enabling technology for 5G. Slicing would allow operators to efficiently run multiple instances of the network over a single infrastructure for serving various applications, use cases, and business services with superior Quality of Service (QoS).

The telecom industry is going through a massive digital transformation with the adoption of ML, AI, feedback-based automation and advanced analytics to handle the

next generation applications and services. AI concepts are not new; the algorithms used by Machine Learning and Deep Learning are being currently implemented in various industries and technology verticals. Existing automation features like Network Function Virtualization (NFV), Software-Defined Networking (SDN) and Network Slicing in 5G are the solutions to generate new sources of revenues, reduce the operation cost incurred by a single core network for diverse services. These features also aim to increase the elasticity and efficiency for scaling new business demands from IoT, Public Safety, Automotive and Healthcare applications.

SDN based architecture is more prone to malicious attacks compared to monolithic core architecture because of the network function virtualization, which creates more entry points available to attackers to infest into the network. With more slices, multiple network configurations and virtual devices, security and privacy concerns in the cellular ecosystem are at a critical juncture, resulting in the need for secure software and methods to build a robust and secure ecosystem. Attacks on the 5G network could have severe consequences on the society in broader aspect. 5G Network Slicing can play a vital role in providing dynamic and flexible security architecture for isolated networks optimized for applications with varying needs from a security perspective by customizing independent firewall configuration(s), security policies along-with slice specific authentication schemes. Our DLNN based DeepSlice and Secure5G model are explained in Chapter 5.

1.5 Research Objectives, Significance and Accomplishments

The main research objectives of this research work are mentioned below:

- Cellular handovers are always often treated as Exponential Arrivals. Utilize the Matrix Exponential (ME) models to capture the multi-dynamic nature of handovers and more accurately characterize the factors impacting handovers.

- Applying the Machine Learning (ML) and Deep Learning (DL) mechanisms to capture multiple real-time radio access side and network key parameter indicators (KPIs) causing a cellular handover which is otherwise not possible manually.

- Design data-driven Deep Learning Neural Network (DLNN) HO models to accurately predict HO and implement all real time network or environmental changes by continuous parameter adjustments.

- Formulating a better representation for disaster failures and recovery processes for a cellular network using the Matrix Exponential (ME) distributions.

- Use Fade Duration Outage Probability (FDOP) to improve network connectivity, reliability and low latency with fractional packet duplication (FPD).

- Attain high degree of Availability using two or more uncorrelated links and only duplicate packets efficiently to not over utilize the limited radio resources applying the proposed Adaptive Fractional Packet Duplication (A-FPD) schemes.

- Orchestrate the Network Slicing Function (NSF) by automating slice prediction, load balancing, and slice failure scenarios using the device key parameters.

- Ensure end-to-end security of our DeepSlice model and avoid any service level interruptions using the proposed Secure5G model.

CHAPTER 2

MATRIX EXPONENTIAL AND DEEP LEARNING NEURAL NETWORK MODELING OF CELLULAR HANDOVERS

2.1 Introduction

This chapter is based on two of our papers published on optimization of cellular handovers, "Handover performance prioritization for public safety and emergency networks" by R. A. Paropkari, C. Beard, and A. Van De Liefvoort, published in the 2017 IEEE 38th Sarnoff Symposium and the second paper "Deep-mobility: A deep learning approach for an efficient and reliable 5G handover," by R. A. Paropkari, A. Thantharate, and C. Beard, published in the 2022 International Conference on Wireless Communications Signal Processing and Networking (WiSPNET). I want to thank my co-authors for their contribution towards this research and their help with writing the papers.

Mobility is the most important aspect in the wireless world and handovers are the way to achieve mobility, with strong expectations that no user equipment will experience any sort of interruption to their services. In the next generation of mobile networks or the 5G networks, a heterogeneous environment of Multiple Radio Access Technology (MRAT) must exist, where all these different networks have to co-exist for all devices of various capabilities. Deployments of small cells and hyper-densification of mobile networks is also another valid reason to make handovers more efficient. In such cases, handovers could be much more frequent and susceptible to drops and insufficient resources.

With the advance standards and mandates, these handovers need to happen with no interruption to the service at any point to maintain the required minimum service functionality.

The ever-growing increase in the number of devices and services has led the mobile industry to grow exponentially in the past few years. Connected devices in future cellular networks are not only limited to mobile handsets and tablets, but belong to a wide range of equipment coming from an overlapping area of eMBB (enhanced mobile broadband), mMTC (massive machine type communication) and URLLC (ultra-reliable low latency communication). Data demands on these devices are huge and the bandwidth requirements are in Gbps due to the video-on-demand services offered today. According to Ericsson's Mobility report from November 2018, the global mobile data traffic is expected to increase eight times between 2017–2023 and video content about five times being offered on over 70% of whole mobile data traffic [3] triggering the eMBB services.

Service providers are forced to provide an infrastructure for serving various applications, use cases, and business with superior Quality of Service (QoS). In some industries like the public safety, emergency, medical, or energy sectors, mobile devices provide great capabilities but must be highly reliable. The current 4G-LTE infrastructure is incapable of meeting high reliability standards of low latency and highspeed user connectivity for URLLC services. Another pool of devices, in the verge of explosion, are the IoT (internet of things) devices catering to the mMTC services of the 5G networks transforming the home, office, city streets, public places, and beyond. These all have a specific QoS requirement and are treated differently in handover (HO) scenarios, if any. Until now, mobility was one of the prime factors to trigger a handover in any given network but several

other parameters like Received Signal Strength (RSSI), Signal to Interference and Noise Ratio (SINR), Fade Duration (FD), Quality of Service (QoS), backhaul connectivity, network congestion, reliability, etc., may also result in a handover. We propose a scheme that will accommodate such parameters to make a combined/intelligent decision towards handover by capturing the multi-dynamic nature of a handover.

We briefly introduce 5G cellular handover concepts and deep learning concepts in Section 2.1, some background work and related work details are in Section 2.2, we explain matrix exponential handover queueing models with numerical results in section 2.3, and in Section 2.4 we present our proposed solution models using RNN/LSTM and the detailed deep learning based Deep-Mobility model with numerical examples and even discuss its application for 5G Mobility use cases. Finally, in Section 2.5 we conclude our work and propose possible future extension.

2.2 Related Work

There has been significant work performed on Matrix Exponential and the PH-type distribution models. Several papers have successfully used such PH-type models to study the performance of wireless networks and some mobility related aspects in cellular networks [4,5]. This PH-type distribution is extremely versatile in network modeling, and the more general ME distributions extend these models [6]. Use of generalized ME distributions also allows techniques both behind and in front of the Laplacian curtain, and additional tools for inverse modeling are now available. Such models have been solved and are numerically well understood in literature. Performance measures such as the connections

blocked, handover connections dropped, and preemption if any, were previously studied and modeled as independent traffic arrivals in heterogeneous networks [7, 8]. The ME approach has also been used to successfully study some other issues regarding hand-off calls using an approximation technique that uses single cell decomposition analysis [9].

The main contribution of our work is in the more detailed modeling of Hand-off traffic, which takes several parameters like SNR, QoS, Fade Duration, Shadowing, etc. into account. Future models will include non-renewal models for the handover traffic, which is characterized as a tuple $(\mathbf{B_1}, \mathbf{L_1})$, where $\mathbf{L_1} = \mathbf{Be'p}$ whenever the process is a renewal. This process generalizes the MAP process in the same way that matrix exponentials generalize the Phase-type distributions [6]. Some studies have considered the numerical computation of stationary distribution for the level dependent quasi-birth-and-death (QBD) process [10] Some theories discuss and compare controlled preemption based queueing schemes to give emergency services and handover traffic priority over other connections [11–13]. Channel holding time (CHT) distributions were derived when cell residence time and call holding time were both assumed to be PH type distributions [10]. Markov based framework was used to define user state and derive a better strategy than a conventional handover optimization technique [14].

Some papers have worked in the construction of a tridiagonal block matrix as we will encounter one in our model, and the inverse transform of the same, to find the normalization constant [4, 15]. In [16] a ME distribution was used to model general arrival and service processes in the analysis and simulation of mission critical publish-subscribe message patterns in a Service Oriented Architecture (SOA). A Random Number Gener-

ator (RNG) was created especially for ME distributions as the authors in [17] strongly agree that ME represents a situation more accurately and descriptively.

As for the best of our knowledge, our work is unique as it is the first to assume multiple parameters from UE and network side to come together to trigger a mobility-based handover decision using a DLNN technique. An eNB pre-selection strategy is proposed in [18] so that high-speed user equipment is primarily taken care by macro eNBs while low-speed UEs or UEs with low quality of service (QoS) requirements are offloaded to the service area of pico eNBs to share the load of macro eNBs. The authors in [19] make use of reinforcement learning (RL) to control the handovers between base-stations using a centralized RL agent. This agent handles the radio measurement reports from the UEs and chooses appropriate handover actions in accordance with the RL framework to maximize a long-term utility. Ericsson mobility report predicts the growth of mobile devices, 5G network connections and the overall data usage in coming years [3]. To overcome handover latency, [20] jointly considers edge and core delays, with a novel cost-effective software-defined ultra-dense framework by dynamically removing state execution times.

The authors in [21] have a game theoretical approach implemented and evaluated for dense small cell heterogeneous networks to validate the enhancement achieved in the proposed method. As for network intelligence, the authors in [22] represented handovers using matrix exponential distributions for public safety and emergency communications, which helps make handover decisions more accurate considering all the different parameters involved in the decision process. To solve the unnecessary HOs and ping-pong, [23] proposes a weighted fuzzy self-optimization (WFSO) approach for the optimization of

the handover control parameters (HCPs) considering SINR, traffic load of serving and target BS, and UE velocity. The 5G Network Slicing concept is fully utilized to manage the network traffic and route the connections to the most appropriate slice using DLNNs and understanding of what the connection demands in [24]. Multiple network and RF parameters are considered before making such an intelligent decision using neural networks. In [25] authors propose a scheme to control Cell Individual Offsets (CIO) and adjust the Hysteresis and TTT autonomously for handover management in order to deploy mobility load balancing (MLB) and MRO independently. [26] developed a Neural Network based 'Secure5G' Network Slicing model to proactively detect and eliminate threats based on incoming connections before they infest the 5G core network.

Authors in [27] contrast Fade Duration Outage Probability (FDOP) based handover requirements with the traditional SINR based handovers methods in cellular systems. The research in [28] evaluates the performance parameter of X2 based handover on one network provider in Cirebon area and optimizes handover parameters using RSRP and RSRQ algorithm. In order to accomplish successful HO, authors in [29] study the impact of HOM, A3offset along with TTT and extend their analysis to different distances from BS for varying UE velocity. Work in [30] proposes an improved MRO algorithm by taking into account TTT to improve MRO algorithm performance in terms of reducing Ping-Pong rate and lower HO failure rate between Femto cell and Macro cell. Authors present system-level architectural changes on both UE and Network elements along with a proposal to modify control signaling as part of Radio Resource Control messages using smartphone battery level in [31] . The work done in [32] shows a direct set-up of the HO

parameters relying on Page Hinkley HO-based decision followed by a closed, iterative loop to further optimize the initial configuration using Simulated Annealing approach.

2.3 Matrix Exponential Based Handovers Modeling

2.3.1 Handover Queueing Models

In cellular systems, we have newly originating connections within a cell and also incoming handover connections from a neighboring cell. Some connections are also handed off outside to neighboring cells, which we are not concerned about for our current study of handover connections. Most modeling of these new and handover arrivals in a cell are treated to be exponential arrivals which don't really capture the multi-dynamic nature of a handover arrival [33]. Even though the major factor for handovers is movement, there is not one single factor in a real-world scenario which triggers a handover. Handover dynamics come from a combination of simple network aspects like the Received Signal Strength (RSSI), Signal to Interference Ratio (SIR), Signal to Noise Ratio (SNR), Fade Duration (FD), Quality of Service (QoS), shadowing, backhaul connectivity, network congestion, reliability or resource availability of a given network, etc. Some other dynamic factors causing a handover are multipath, Doppler spread, or velocity. Even the slightest change in the location in a dense cellular network may cause a user to handover to a nearby or a neighboring cell.

The single-parameter memory-less exponential distribution is not at all sufficient to capture these effects, instead, all of these factors can very well be analyzed and studied if we are to represent a handover arrival as a Matrix Exponential (ME) distribution (details and derivation in Section II.B). This can provide an opportunity for looking at each

of the factors and parameters in the previous paragraph individually and also how they all work together to cause a handover. Also, use of an ME can help us calculate the blocking probabilities much more accurately than precisely using any individual exponential parameter. Connection drop probability is one important parameter that is used to measure the Quality of Service by any cellular service provider. Our ME representation can lead us close to the most accurate values. Various parameters captured using an ME can not only help us study their impact on the handover process but as well help design a very efficient handover algorithm which would otherwise be difficult with only one of the parameters under consideration.

In the case of ultra-reliable communication and critical or emergency services, we must always take precautionary measures to make sure that such a connection gets minimum required resources, and if any handover action is to be taken, is done well before the service deteriorates. For example, if SNR is the only handoff trigger factor, then maybe we have an acceptable SNR in an area, but the radio resources available are not enough and so we get low throughput in spite of having good SNR. If there's an overlapping cell in that region which can provide greater throughput at a slightly lower SNR then the user needs to be handed off to this new cell.

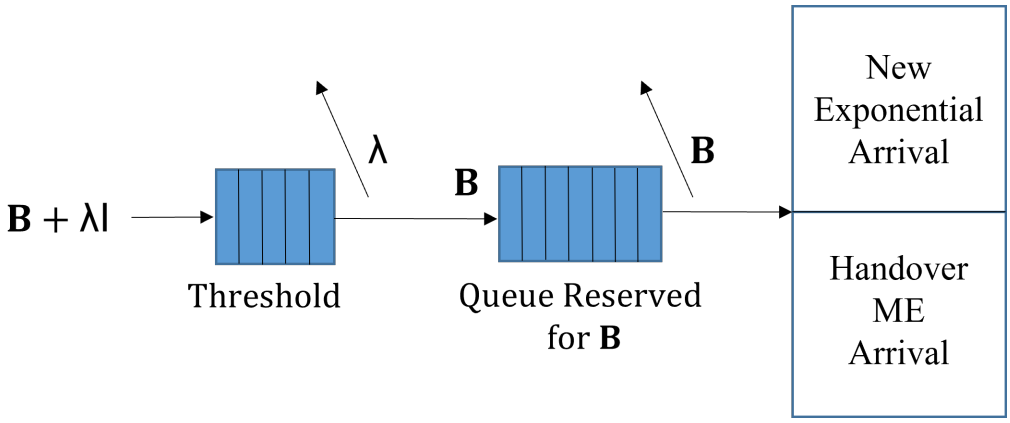

Figure 1: General Representation of our Queueing Model

2.3.2 Matrix Exponentials

Matrix Exponential (ME) distributions are the probability distributions whose distribution functions and densities are defined as [34].

$$F(t) = 1 - \mathbf{p}.e^{-\mathbf{B}t}\mathbf{e}' \quad \text{for } t \geq 0 \tag{2.1}$$

$$f(t) = \mathbf{p}.e^{-\mathbf{B}t}\mathbf{B}\mathbf{e}' \quad \text{for } t \geq 0 \tag{2.2}$$

In this work, **p** is a row vector of appropriate size and **e** is a row vector with all 1s. So $\mathbf{e}'$ is a column vector and a transpose of **e**. The Matrix Exponential (ME) distribution carries with it the simplicity of capturing Markov chain analysis but transition rates are modeled by matrices that can approximate virtually any distribution with an appropriately sized and formed matrix. For illustration purposes we show our handover representation as more of a Phase Type (PH) distribution, but an ME distribution is much more general and only need to have a rational Laplace Transform. They are more general and flexible to work with.

As stated previously, ME representation can very well capture all the factors that

will lead to a handover. It has proven its credibility and advantages in many other fields in the society. Some of the medical institutions and hospitals make use of ME queueing models to illustrate the resources that would be available at any given point in time.

2.3.3 Proposed Queueing Model

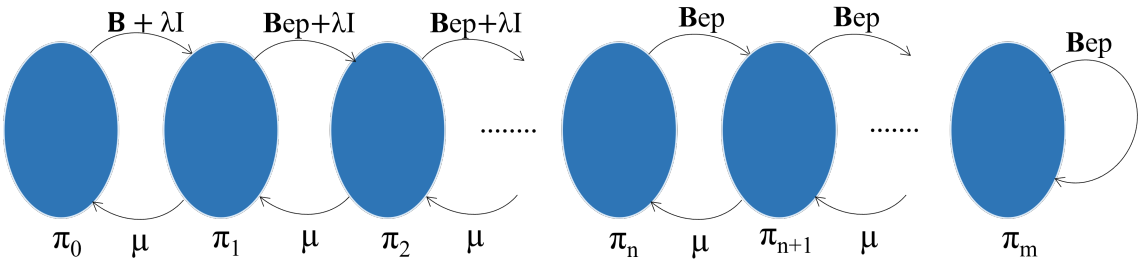

Figure 2: Queueing Model of our System

In our model of a given cellular network, we treat any newly originating arrival as exponential (λ) whereas an incoming handover arrival is a Matrix Exponential (ME – represented as a **B** matrix) in the system. We have a system, where in any given cell, we have limited resources available, that we consider as having a total number of m available channels. Now our model accepts both exponentials as well as the ME arrivals into the system. And we have a single server that serves the arrivals and everything else can be queued in our main queue. As an example, we use a queue size of m=10. However, the originating arrivals can only be queued until the queue size reaches buffer size n=4, which is when we stop receiving the exponential originating arrivals. Then for the remaining states in the queue, only the ME handover arrivals are accepted in the system and queued until the server is available to serve. The overall queue size is also limited in our case which limits to the number of ME handover call coming into the system. The general

prioritization concept is shown in Fig. 1, and queueing model Markov chain is shown above in Fig. 2. From this basic formulation, we can readily extend the model to use different queue sizes, priority thresholds, and number of servers (channels).

Fig. 3 shows how we compute **B** matrix representing the ME arrival handover process. Each large oval corresponds to the ovals in Fig. 2. Inside we can have many possible transitions that create conditions that affect handovers. For example, as a user moves, the combination of decreasing RSSI from the home base station and the increasing interference of a nearby base station both combine for a real need for a handover. Only those arrows that leave the large left oval represent a handover trigger. Based on the state transitions within each of the two large ovals, we can compute a **B** matrix as shown below. The **Be** process captures a handover due to all of the effects of changes in state, and the **p** vector randomly determines which states begin the process for the next handoff. The **p** vector can be considered as a beginning or starting vector of a new state that a system enters and waits for another handoff to be triggered in a similar fashion. This is under the assumption that our handover process is renewal that is the next handover process starts all over again in a new state after a handover is triggered. This is not entirely accurate because the next mobiles that will become candidates for handover were already moving towards handover. But this depth of accuracy is the subject for our next work, and the **Bep** renewal representation accurately captures many handover dynamics.

$$\mathbf{B}_{\mathrm{B}} = \begin{pmatrix} A_1 + a_1 & -a_1 & 0 & 0 \\ -b_1 & A_2 + a_2 + b_1 & -a_2 & 0 \\ 0 & -b_2 & A_3 + a_3 + b_2 & -a_3 \\ 0 & 0 & (\mu_{2c} + \mu_{2e}) & -\mu_{2c} \\ 0 & 0 & -b_3 & A_4 + b_3 \end{pmatrix} \tag{2.3}$$

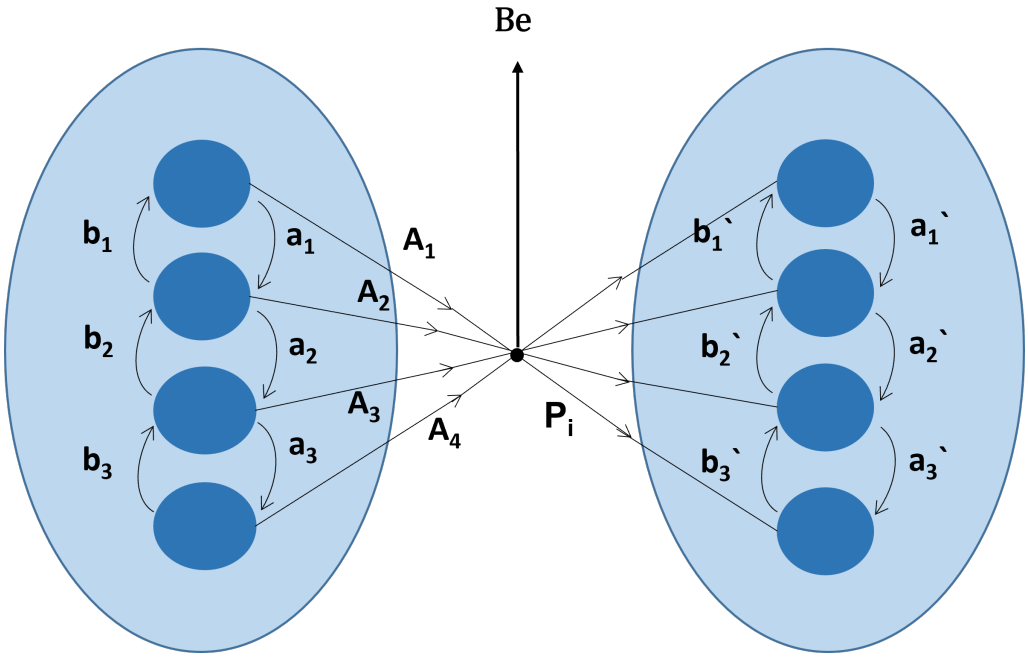

Figure 3: **B** Matrix formation for a Handover Representation

We have already mentioned a few parameters or factors that can cause a handover to trigger. We have tried to model only four such parameters just so we can see how they, either individually or collectively, cause a handover to happen and this also helps us to keep our model as simple as possible. As you can see from the **B** matrix model that the initial state has four internal states (representing four individual parameters) and transitions can come out of their states to cause a handover (shown as A1, A2, A3, A4) and then enter a new state. This new state can actually start in any of these four internal states with any values based on the new cell conditions for these same four parameters. The **p** is a row vector that guides the beginning of the new state and $\mathbf{e}'$ is a column vector of 1's. Also, when in the initial state, these parameters can slightly change and cause transitions to another internal state (shown as a1, a2, a3) causing a combined effect of at

least two or more parameters to cause a handover. There exists some probability also that the internal states can go in either direction, even back to the original (shown as b1, b2, b3) in Fig. 3 above.

New arrivals (which arrive according to a Poisson process) are accepted into the system only if n (or fewer) costumers are in the system, and are otherwise blocked-and-cleared. Similarly, handoff arrivals are accepted into the system if m (or fewer) customers are in the system, and are otherwise blocked-and-cleared as you may see in the last state of our model. Based on our system model, until the queue size of n is reached, we are accepting both types of arrivals and the representation in terms of balance equations is given as follows. For the first n states (in our example) exponentially arriving originating arrivals of rate λ are accepted and queued, so the arrival rate is $(\mathbf{B} + \lambda\mathbf{I})$. All arrivals of both types are assumed to be served at an exponential rate μ. In the future, this process could also be replaced by ME matrices.

$$\begin{aligned} \boldsymbol{\pi}_0\mathbf{B_1} &= \boldsymbol{\pi}_1\mu \\ \boldsymbol{\pi}_i\left(\mathbf{B_1} + \mu\mathbf{I}\right) &= \boldsymbol{\pi}_{\mathbf{i-1}}\mathbf{L_1} + \boldsymbol{\pi}_{\mathbf{i+1}}\mu \qquad \text{for } 1 \leq i \leq n \end{aligned} \tag{2.4}$$

where $\mathbf{B_1} = (\mathbf{B} + \mu\mathbf{I})$ and $\mathbf{L_1} = \mathbf{B}\mathbf{e}'\mathbf{p} + \lambda\mathbf{I})$

Now for the remaining states between n and m-1, where only the ME handover arrivals are accepted, the balance equations are somewhat simpler, based on our model.

$$\begin{aligned} \boldsymbol{\pi}_j\left(\mathbf{B} + \mu\mathbf{I}\right) &= \boldsymbol{\pi}_{\mathbf{j-1}}\mathbf{B}\mathbf{e}'\mathbf{p} + \boldsymbol{\pi}_{\mathbf{j+1}}\mu \qquad \text{for } n+1 \leq j \leq m-1 \\ \boldsymbol{\pi}_m\mu &= \boldsymbol{\pi}_{m-1}\mathbf{B}\mathbf{e}'\mathbf{p} \end{aligned} \tag{2.5}$$

These equations are solved numerically, but we would like to add that the station-

ary distribution satisfies the following equation for $n+1 \leq j \leq m-1$

$$\boldsymbol{\pi}_{j+1} = \boldsymbol{\pi}_j \boldsymbol{U} \tag{2.6}$$

and thus

$$\boldsymbol{\pi}_{n+k} = \boldsymbol{\pi}_{\mathbf{n}} U^{k\text{-}1} \tag{2.7}$$

where

$$\boldsymbol{U} = \frac{(\mathbf{B} + \boldsymbol{\mu}\mathbf{I} - \mu \mathbf{e}'\mathbf{p})^{-1}}{\mu} \tag{2.8}$$

The above equations stand true as we consider the handover traffic being a renewal process. And in this case, the value of $\boldsymbol{\pi}_{\mathbf{n+k}}$ is known exactly and symbolically (up to normalization), and is a matrix geometric with the U matrix as shown above. Here n is the last state when the queue size reaches the threshold of accepting originating arrivals and m is the last state in our system where no more handover arrivals are accepted and our queue is full. We define our threshold n as the threshold queue size up to which originating connections would be accepted. Therefore, we can define the following blocking probabilities. These are also known as the time blocking probabilities, which indeed are for the fraction of the time that the system is in a state where a ME handover connection could not be accepted.

$$P_{block,orig} = \boldsymbol{\pi}_{\mathbf{4}} + ... + \boldsymbol{\pi}_{\mathbf{10}} = \sum_{i=n}^{m} \boldsymbol{\pi}_{\boldsymbol{i}} \tag{2.9}$$

$$P_{block,HO} = \boldsymbol{\pi}_{\boldsymbol{m}} \tag{2.10}$$

2.3.4 Numerical Examples

We consider a simple handover scenario where we express a newly originating cellular connection as an exponential arrival represented by lambda (λ) and the incoming handover connection as a matrix exponential represented by a 2-by-2 **B** matrix. Later we could easily use a larger **B** matrix to even more accurately model a system. We have considered Hyper-exponential and Erlang-2 phase-type models for the handover connection (even though matrix exponential can model beyond phase types). We modeled these arrivals with a queue size of m = 10 and our initial buffer threshold is n = 4 where the newly originating connections as well as the incoming handover connections are both accepted. Once the buffer is full to the threshold, we start rejecting the exponential originating connections but only accept the ME handover connections, until our queue size of m = 10 is full. No connections will be accepted then. We calculate the probabilities at each state, and from them find the blocking probabilities for the newly originating arrivals as well as the dropping probabilities for the handover arrivals. We have also solved these numerically. And note that for the coefficient of variation, C^2=$E(X^2)$/$(E(X))^2$, we have C^2<1 for Erlang-2 handover arrivals, C^2=1 for exponential and C^2>1 for hyper-exponential handover arrivals.

The following plots help us analyze how an ME representation of handover arrival can be more beneficial and efficient to design modern handover algorithms. These will make next generation 5G networks perform acceptably for emergency and critical communications where dropping a handover connection is not permissible. For the first two plots in Fig. 4 and Fig. 5, we fix our queue size to m = 10 and vary the threshold from n =

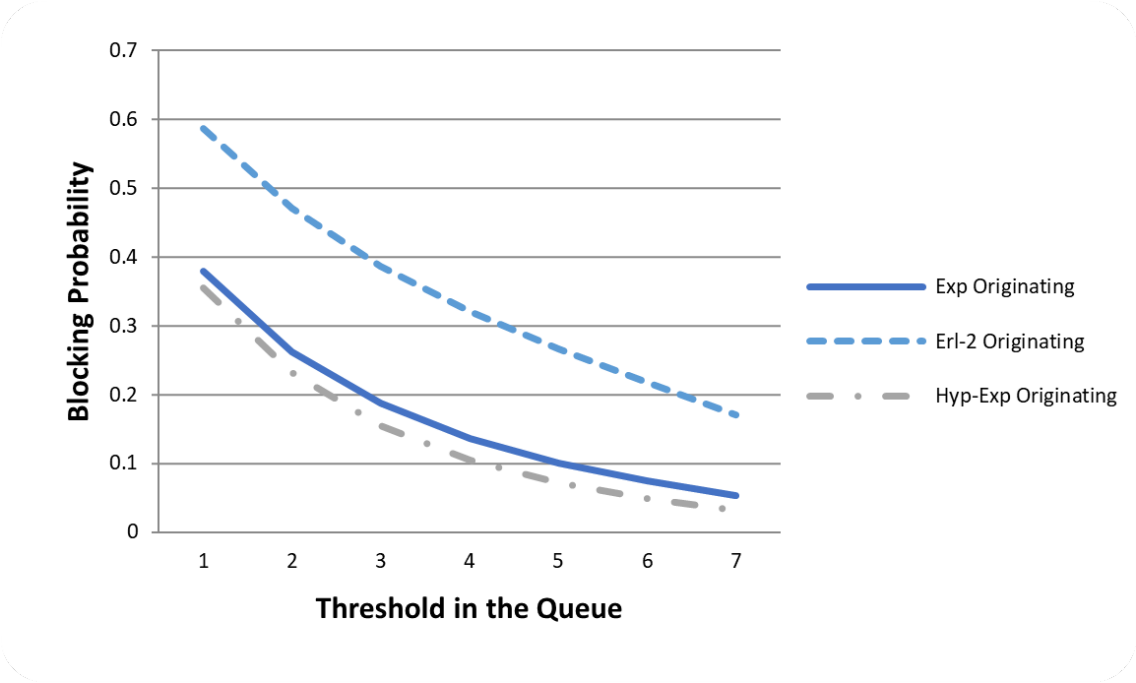

Figure 4: Varying the Threshold for Originating Arrivals for $\rho < 1$

2 to n = 8 and plot the blocking probabilities for the newly originating connections as well as for handover connections. Originating and handover arrivals were plotted separately for exponential arrival and ME arrivals. We show the exponential plot where originating and handover arrivals were both exponential in the same queue.

The originating arrival rate is set to 0.3 and we model the handover arrivals as hyper exponential as well as Erlang-2. But we make sure to keep the combined arrival rate fixed at 0.7 in both the cases. Our service rate is exponentially distributed with mean 0.9 which gives us $\rho < 1$ and it is clear from the Fig. 4 that increase in threshold will significantly drop the blocking probabilities for newly originating arrivals. Fig. 5 indicates how the handover arrivals will see more blocking as originating connections will also be

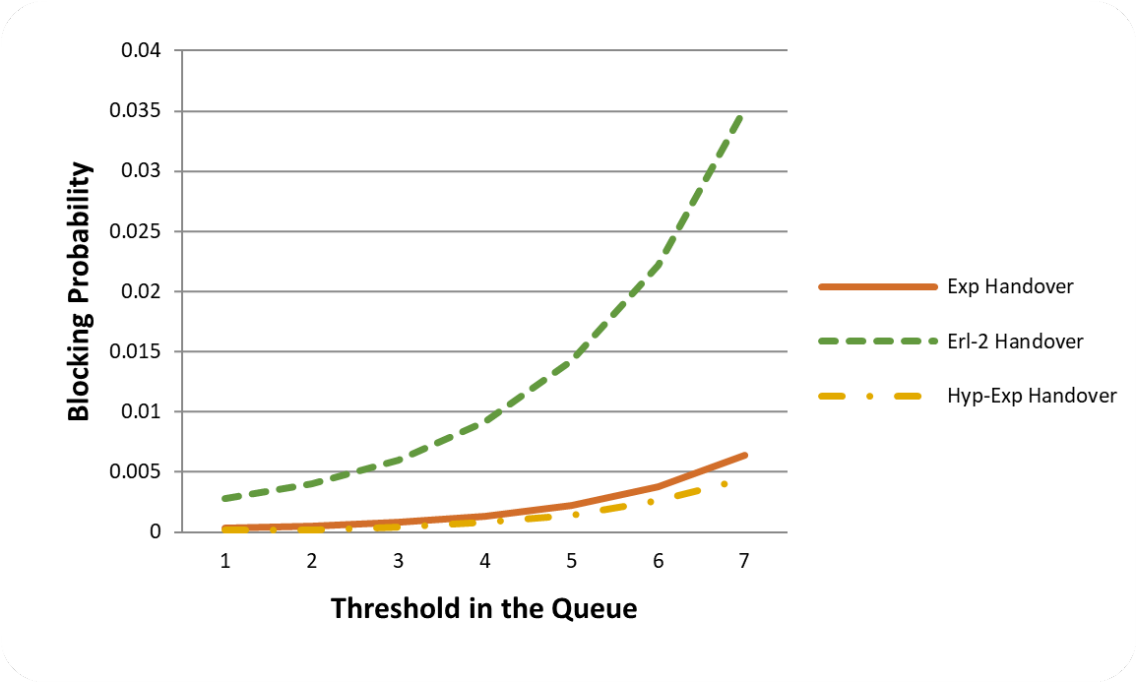

Figure 5: Varying the Threshold for Handover Arrivals for $\rho < 1$

accepted with increasing threshold. The most important observation, however, is that the blocking probabilities are drastically different depending on the type of handover matrix exponential arrival.

We now consider $\rho > 1$, again we see a significant difference from the exponential arrivals case depending on the arrival process for the handovers. In order to get $\rho > 1$ as shown in Fig. 6 and Fig. 7, we reduce the service rate to 0.4 in our queue. The drop in blocking probability for originating arrivals is now more gradual and handover arrival dropping probability also increases slightly with the threshold. Fig. 6 shows high blocking probabilities for originating arrivals as the threshold is low earlier. But as we increase the threshold, blocking probability drops gradually. Fig. 7 shows that handover arrival

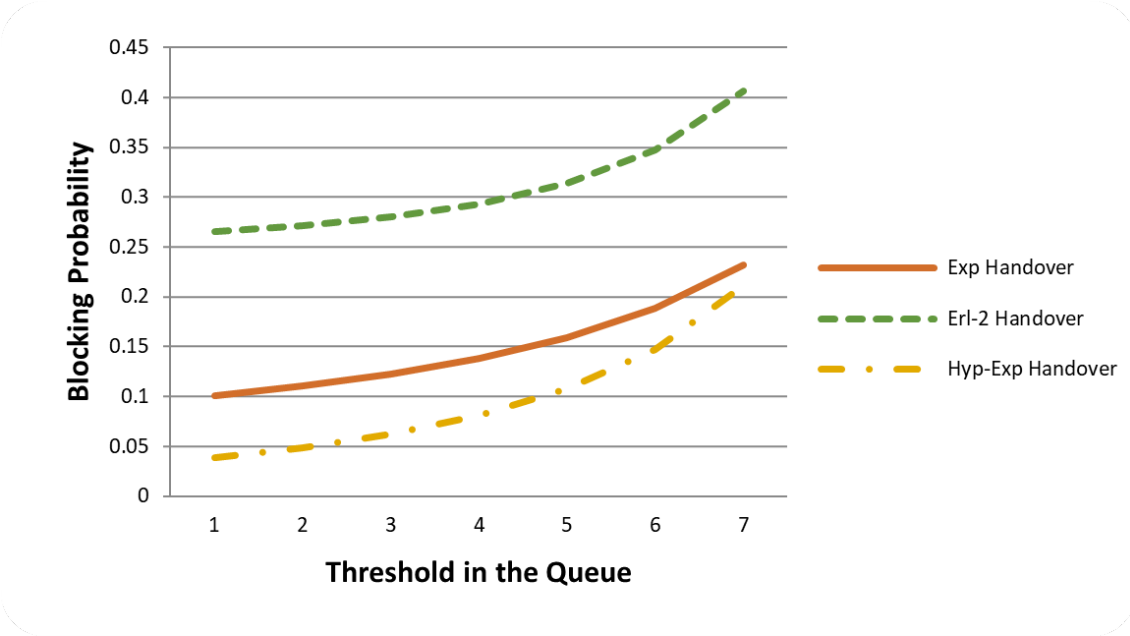

Figure 6: Varying the Threshold for Originating Arrivals for $\rho > 1$

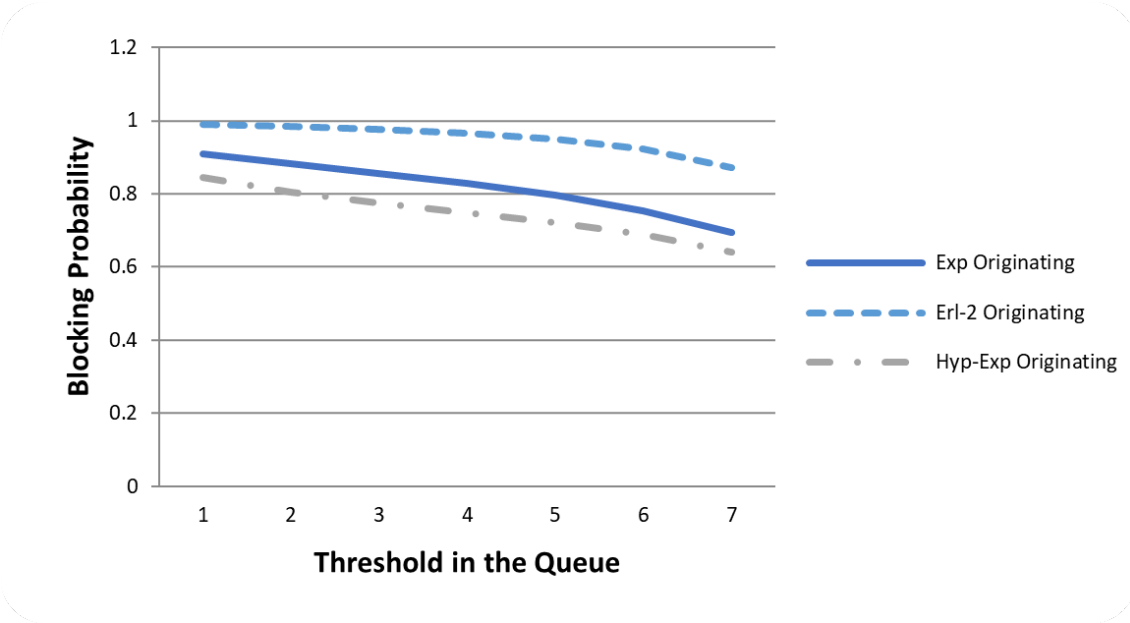

Figure 7: Varying the Threshold for Handover Arrivals for $\rho > 1$

dropping probabilities are higher for a lower service rate and increase with the threshold.

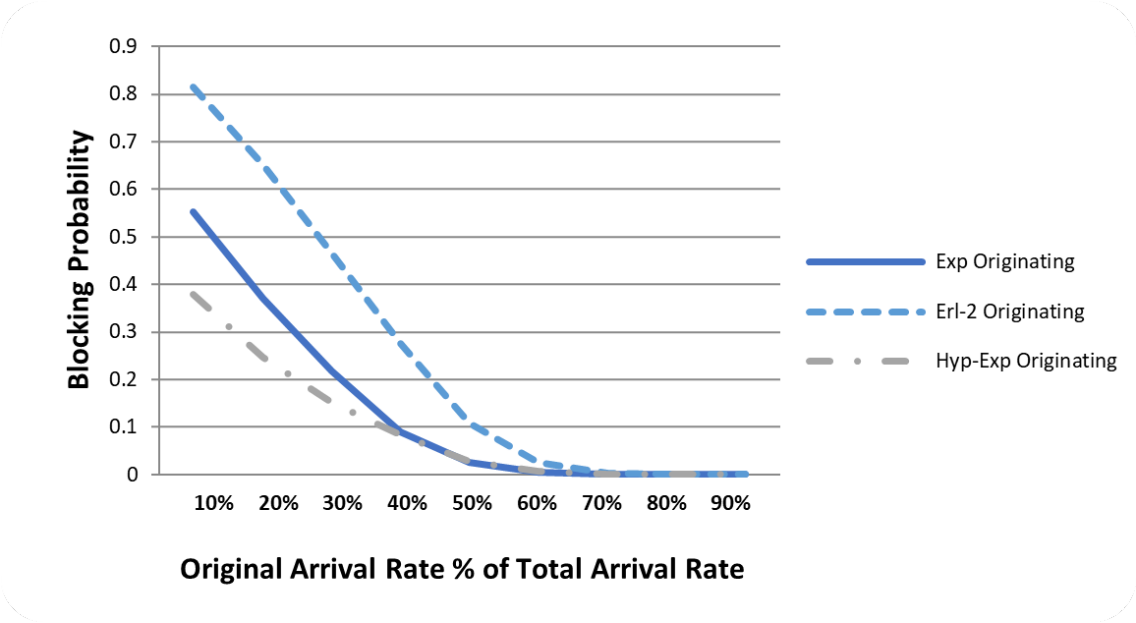

Figure 8: Blocking Probability for Originating Traffic versus Percentage of Total Arrival Rate

Now we consider a scenario where the majority of the arrivals can either be the newly originating connections or the handover connections. Fig. 8 and Fig. 9 show how the blocking probabilities are affected as we vary the newly originating arrivals. Starting with 10% of the overall system arrivals to be originating, we go up to 90% of all those arrivals being originating arrivals. We observe from the plots that as more and more of the arrivals are from originating connections, they get served with lower blocking probabilities. Handover arrivals have better chances of being served in all the three scenarios be it the exponential, hyper exponential or Erlang-2 as shown in Fig 6b. The arrival rate for the newly originating connections (λ_{orig}) on X-axis is varied from being 10% to 90% of the total system arrival rate (λ_{total}) which is the combination of the originating ar-

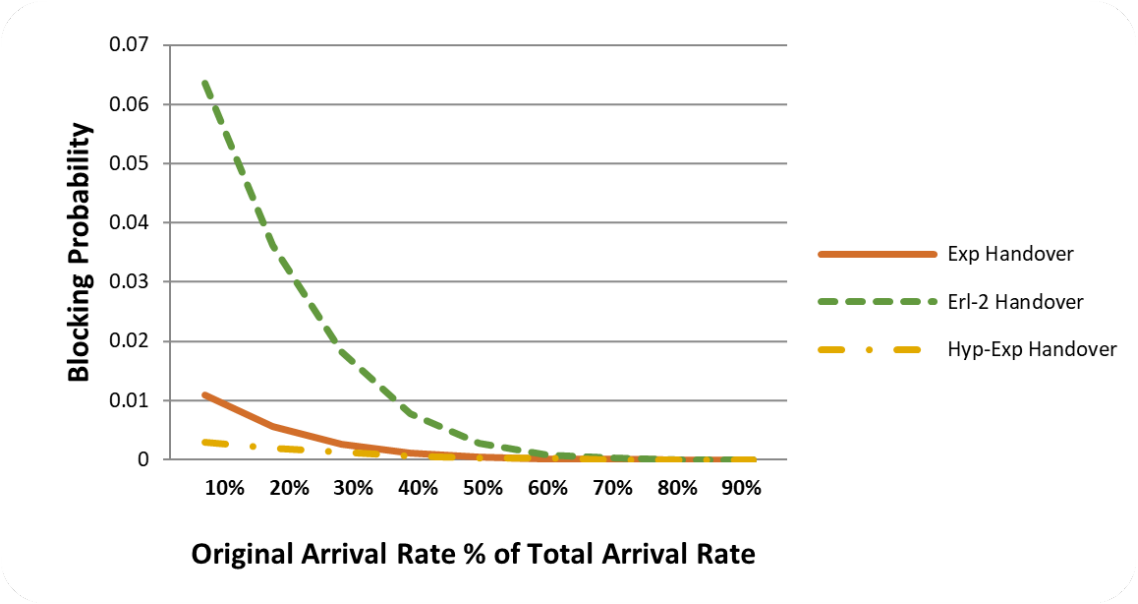

Figure 9: Blocking Probability for Handover Traffic versus Percentage of Total Arrival Rate

rivals and handover arrivals (λ_{ho}). Our queue size of m=10 and threshold of n=4 remains unchanged. Service rate is also fixed at 0.9 and we make sure total arrival rate (λ_{total}) is always the same, no matter how the arrival changes for individual connections, be it newly originating or the handover connection. Eventually handovers will drop too as the queue becomes full. We show the effect of the total queue size. A lower service rate can add delay to processing connections which may also result in dropping some of the handover connections. Our system is a finite system and hence can take some delay even if the service rate is lower. We vary the queue size to see how the state probabilities are affected.

Queue sizes of m = 10, m = 15, and m = 20 are considered but the threshold is fixed at n = 4. The newly originating arrival rate is 0.3 and handover arrival rate is 0.2 while the

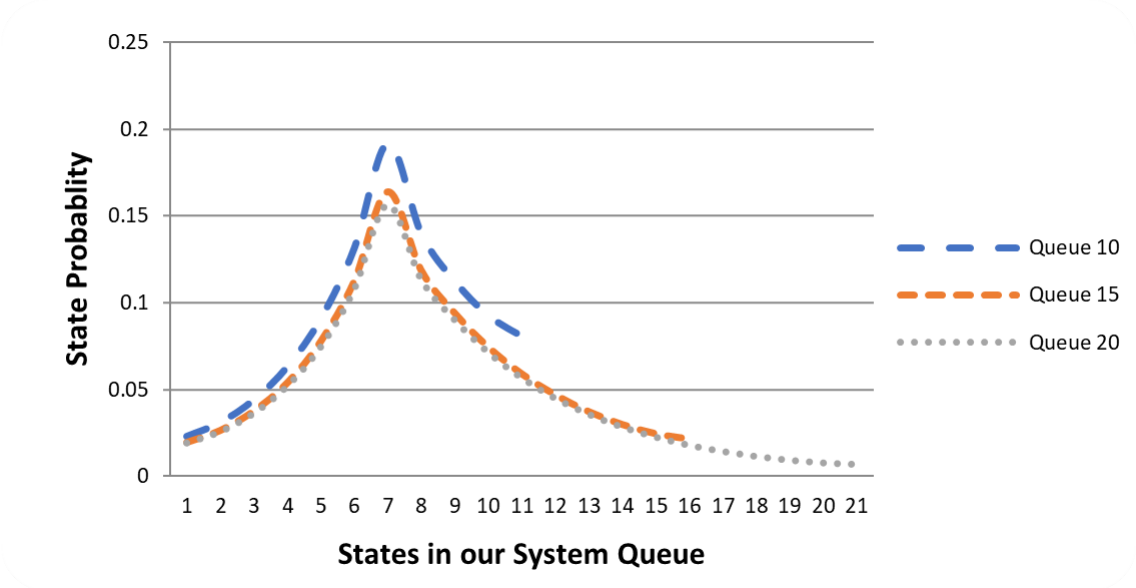

Figure 10: Varying the Queue Size for Hyper-Exponential Handover Model

service rate is fixed at 0.4 this time. This produces a $\rho < 1$. Fig. 10 shows results of the hyper-exponential handover arrival. For the given rates, queue size and buffer, the system has probabilities of serving more connections well before the queue gets full at size of m=10. So increasing the queue size to m = 15 or m = 20 does not matter as connections arriving will be queued and processed as per the resource availability. We have not shown the results for Erlang-2 arrivals since the effect of queue size has more of an influence with the hyper-exponential model.

We have also plotted the strong changes in blocking probability for the hyper-exponentials in range $0.5 \leq C^2 \leq 3$. Fig. 11 shows for handover connections that the blocking probability of 0.00132 for exponential arrivals when C^2=1, drops to 0.00026 (5% of exponential) for C^2=1/2 and up to 0.0061 (over 5 times exponential) for C^2=3. Effects on blocking of originating connections go from 0.234 for exponential down to 0.208 and up

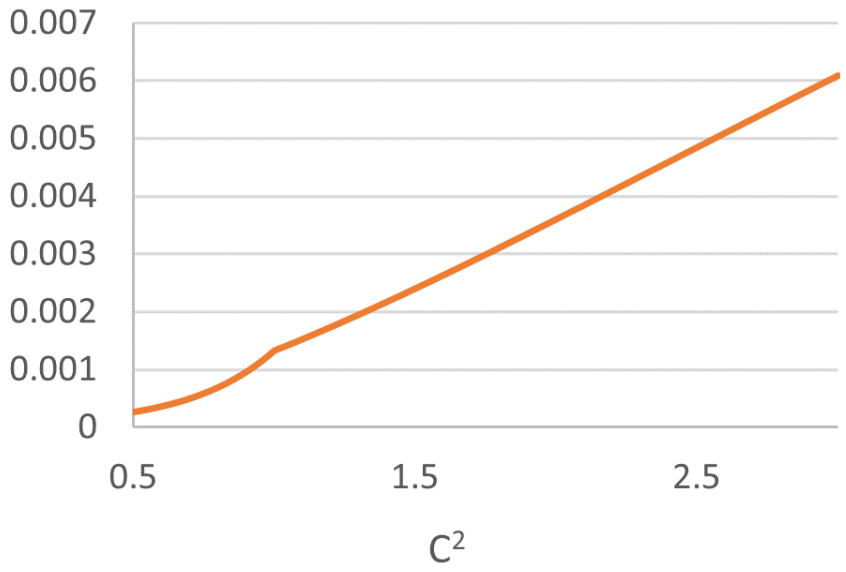

Figure 11: Blocking versus C^2 for Handover Arrivals

to 0.282, as seen in Fig. 12.

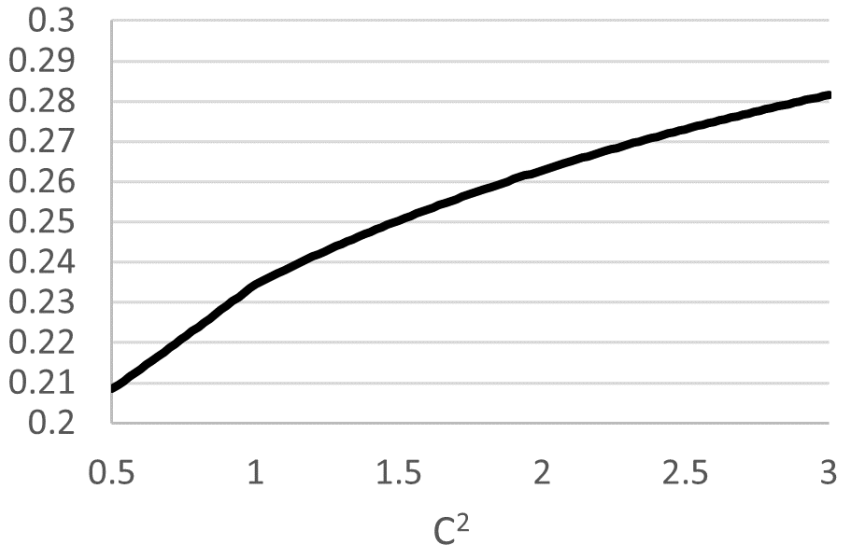

Figure 12: Blocking versus C^2 for Originating Arrivals

2.4 DLNN Based Next Generation Cellular Handovers

Network controlled and UE assisted hard HO procedures are adopted in 3GPP LTE-Advanced. Handover Margin (HOM), Time-To-Trigger (TTT) and A3offset are the major parameters based on which the entering condition of A3event is initiated. The existing HO mechanisms are damaging the 5G latency and availability requirements as about 70% of the current cellular HO are still depending on the RSRP (signal power) and/or the RSRQ (signal quality) as a single parameter considered to trigger a handover. With recent advancements, the management of cellular handovers has become more complex. Networks have become heterogeneous (HetNets) with multiple technologies achieving the same goal of providing endless connectivity to a user. With the on-going increase in the

complexity of the network, manual tuning is very time consuming and very much more prone to errors than before. In addition, RF and channel conditions may change due to several environmental and surrounding factors. Human intervention at every critical step isn't enough, so SONs are gaining importance.

Service providers are implementing some of the SON use cases such as the Mobility Robustness Optimization (MRO) introduced in 3GPP Release 9 [35]. However, the conventional MRO algorithm tunes handover parameters based on counts of handover failures and handover events; and only makes use of the hysteresis to improve HO performance. Unlike in the past, multiple network parameters must be tuned and synchronized to achieve certain goals. 3GPP had also proposed the Automatic Neighbor Relation (ANR) tool for early LTE deployments which is still widely used in the industry. Many HO research areas propose modifying the HOM and TTT in order to make handover decisions precise and accurate for different environments. Achieving load balancing has been an important goal of many academic research areas, and service providers implemented similar solutions like Wi-Fi Offloading, traffic balance using the Femto Access Points (FAP), etc.

Handovers need to keep up a balance in order to avoid the ping pong effect and also too much loading of any one specific cell. With the current and future networks consisting of massive small cell deployment, especially to overcome the uncertainty of millimeter wave (mm-wave) communications, it is important to minimize call drops during HO and avoid redundant and unnecessary HO. A small yet temporary obstacle between a UE and any mm-wave gNB can impact connectivity and trigger a HO even without UE mobility.

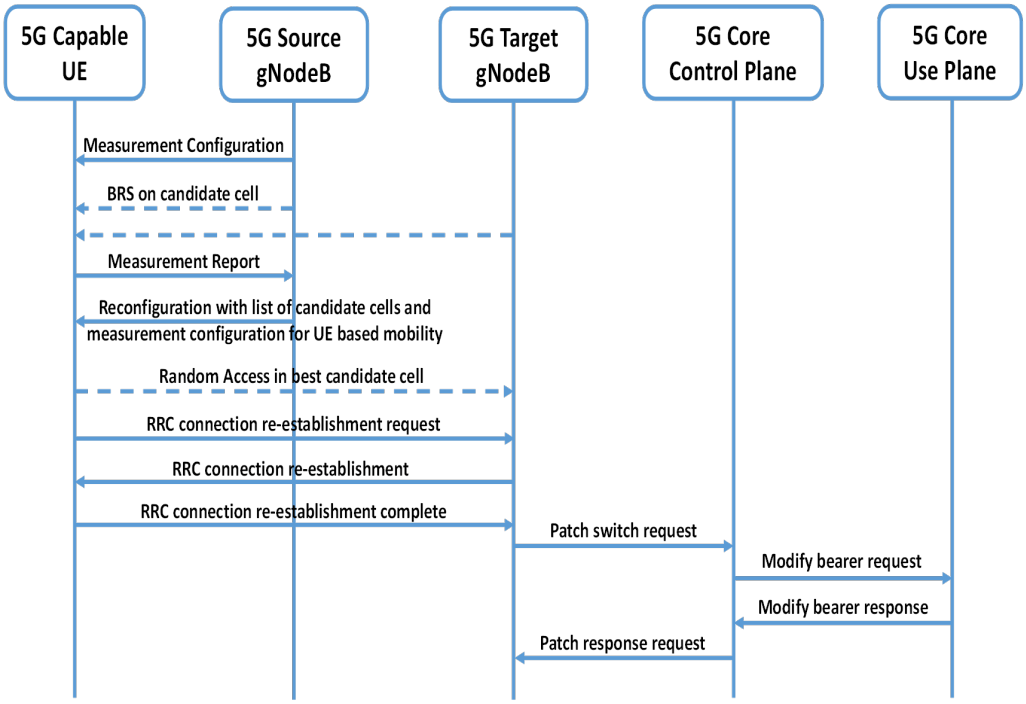

Figure 13: UE/Network Initiated Handover Procedure

The increasing probability of HOs may cause HO failure (HOF) or HO ping-pong which degrades the system performance. Use of SONs are to raise these automation standards and provide the required balance. If more SONs are running with the same goal on different objects, then there's a conflict of interest and that network parameter can be altered wrongly. Certain cell (C-1) might be overloaded and needs some users moved to a neighbor (C-2) and SON-1 takes appropriate action and moves 20% users from C-1 to C-2 by adjusting power levels of C-2. At the same time, SON-2 is making sure all users get a RSRP of a certain value which C-1 can provide at this time and not C-2. So, it's trying to move users back to C-1 causing an overload and ping pong effect.

The main reason for adjusting the same parameter is that not all parameters are

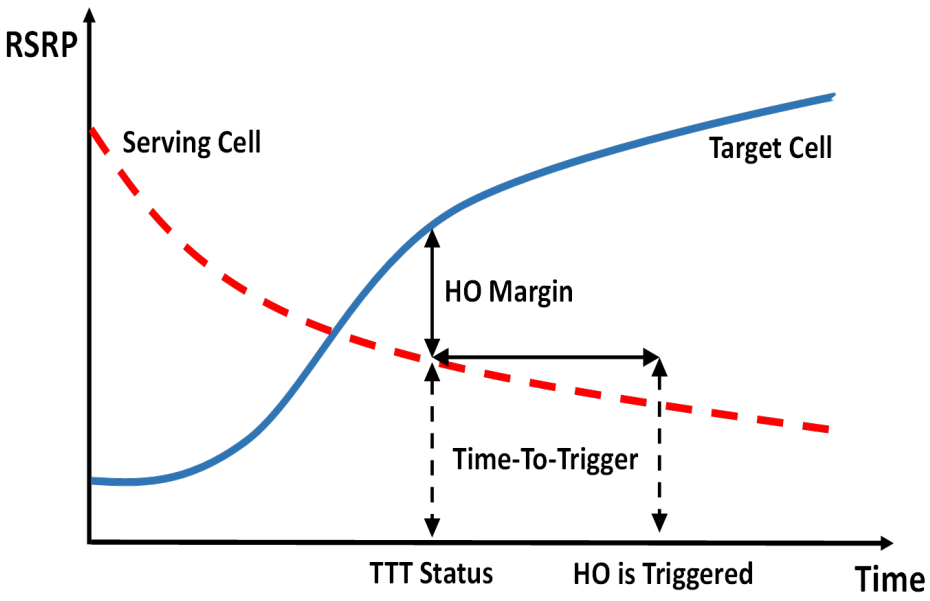

Figure 14: Handover Triggering Decision making Parameters

equally attractive from the optimization viewpoint. For example, some 3GPP mobility parameters, such as the TTT and the Hysteresis, are not defined on a per cell-basis or adjacency-basis, which is a severe limitation in mobile networks due to the variant nature of the radio environment. On the contrary, the HOM is a mobility parameter defined per adjacency-basis that provides greater optimization capabilities. In short, multiple parameters need to be considered at the same time to make an educated decision. Fig. 13 shows the basic handover message exchange which can either be initiated by the UE or by the network. Fig. 14 shows HO requirements such as the HOM or the Hysteresis which is the minimum threshold RSRP value between the serving cell and the target cell power levels. If this value is reached, network begins to count the TTT and if this value is maintained for a minimum of pre-defined TTT, a HO will be triggered. All the above

mentioned justifies the need to involve more UE and network related parameters and network-based triggers towards a HO decision model.

Currently, the base station neighbor list is populated using the ANR functionality of the SONs, but not on a real-time basis. Service providers are looking for a non-static and dynamic way of doing this on instantaneous basis. So, we have a list of neighbors, but then additional parameters are required to figure out which one to connect to if need be. Rather than only considering power levels, HO decision can be made using multiple other parameters like the bandwidth capacity of a gNB, current user load, backhaul capacity, future maintenance activities in the database, RF channel stability, modulation and coding schemes (MCS), dual connectivity (DC), etc.

We create a dataset that contains some information from the UE and some from other network resources. Our dataset consists of the UE type used, supported technology, time and day of the connection, RSRP, RSRQ of serving as well as some (3-4) neighboring cell sites, RF channel conditions based on MCS, available channel bandwidth, backhaul capacity, alarm status on cell sites, maintenance tickets if any, etc. Our DLNN learns from this information and will help identify the following: Too many HOs and when/where, neighbor list needs to be updated, power levels of each eNB at every location, RF conditions, behavior of the UE, movement and velocity of the UE, routine mobility patterns of the UE, power levels of the eNB in UL, UE bandwidth demands, applications being used and other service parameters the from network and RF perspective. Since service providers are deploying dense networks using pico, femto, UE relays, DAS systems, cellular and Wi-Fi hotspots, they are working harder to find a trade-off between

cost, coverage and complexity of network and handover management.

2.4.1 Proposed Solution Models - RNN and LSTM

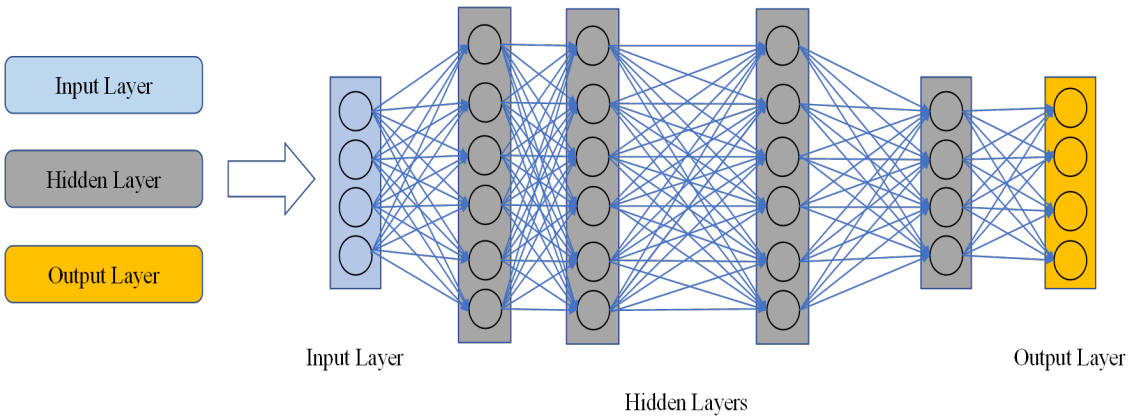

Figure 15: General Deep Learning Neural Network

In today's day and age, prediction has become very intuitive and common on most platforms. However, the prediction problem (and algorithms) is still one of the hardest in data science industry including a wide variety, as in predicting sales, stock markets, speech recognition, sequential prediction on what you will type next while searching, or even some sort of guess work by Alexa/Google/Siri assistants, etc. With NNs being in buzz today, and also practically implemented in many forms, the use of LSTMs has proven to be the most effective solution so far. Recent studies have proven some other less computational and yet equally effective solutions, but LSTMs stay in business. Here we consider RNNs and LSTMs, for reasons listed below.

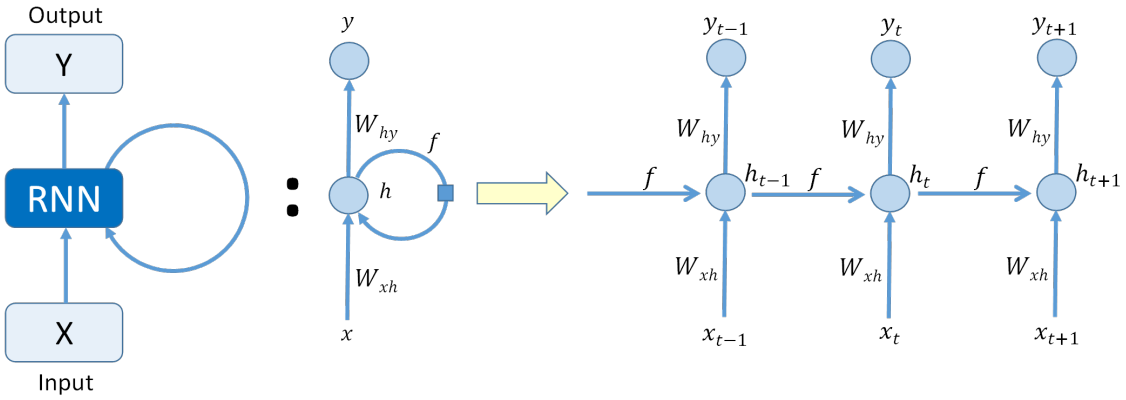

Figure 16: A typical Recurrent Neural Network Model

2.4.1.1 Recurrent Neural Networks (RNNs)

In any conventional feed forward network, all variables are treated independently, meaning any prediction done will not take into account previous output results or any data from a day before. Therefore, RNNs come in play to achieve the time dependency. Fig. 16, shows what a typical RNN will look like and what it mathematically means when expanded. RNNs will take into consideration an event from the immediate previous state and will let the present operation be influenced from earlier results. RNNs have several limitations, and for a continuous time dependent string of data, they tend to not completely understand the context behind an input. They are good, especially when dealing with short term dependencies and universal inputs. So, to go back in the past and relate to something from that far, to help make a decision today, is out of scope of the plain RNN networks and so LSTMs were introduced.

2.4.1.2 Long Short-Term Memory Neural Networks (LSTM)

RNNs often tend to fail in extracting the most appropriate context from the long data feed and this is typically directly related to the vanishing gradient problem in neural networks. When we look at how a network learns, any weights applied are actually a cross product of the learning rate, error from earlier layers and the input to this particular layer. The previous layer error is a product of all previous errors which are theoretically small values. When an activation function is applied to any of the layers, for example a sigmoid function, even smaller values of the derivatives of errors get multiplied several times, resulting in an infinitesimal small value propagated backwards to the earlier layers in the network. And thus, the gradient almost vanishes as we go back toward the earlier layers in the network. This is what happens in standard RNNs where immediate information is available for a short period, but with large datasets, LSTMs are the way to achieve accuracy and reliability.

Unlike the RNNs which apply a single function to transform the whole information, LSTMs follow a slightly different approach of splitting up the information into relevant and non-relevant ones. As shown below in Fig. 17, this is achieved with some simple math functions of additions and multiplications, which allows for more control over the flow and mixing of inputs as per trained weights. In addition, LSTM information flows through 'state', this state is the horizontal line on the top that allows for basic mathematical function to alter the output state. Standard activation functions like the sigmoid (output between 1 and 0) and tanh (output between +1 and -1) are also used to take limited information to the next states.

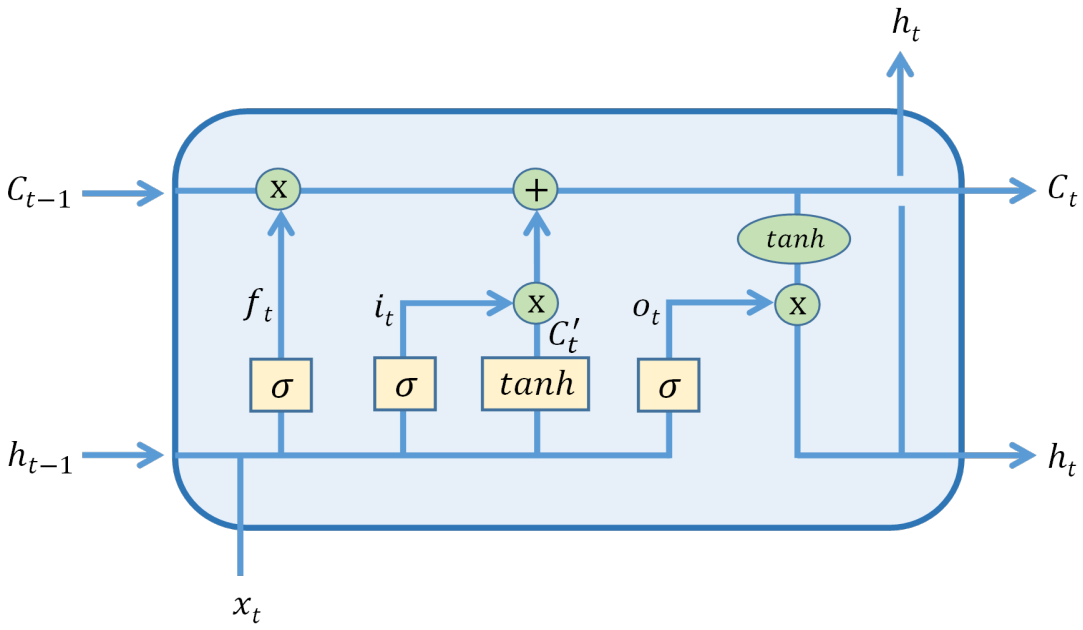

Figure 17: Standard Long Short-Term Memory Neural Network

The LSTMs have the ability to remove or add information to the cell state, carefully regulated by structures called gates. The first and the left most is the 'forget gate' (output denoted by f_t), which applies a sigmoid σ function to previous h_{t-1} and x_t to allow whatever information to take in, 1 meaning take all and a 0 forgets all. The second, and central portion, is a combination of two small functions: first is the input gate in conjunction with a sigmoid function (output denoted by i_t), and second is the $\tanh$ function to create a new vector C'_t that gets added to the state above. This will give us the new state C_t based on the applied functions and the previous state C_{t-1} as shown in below derivations. Finally, we get to our output which will be based on our cell state, but will be a filtered version (denoted by h_t). First, we run a sigmoid layer which decides what parts of the cell state we're going to output (denoted by o_t). Then, we put the cell state through

$\tanh$ and multiply it by the output of the sigmoid gate, so that we only output the parts we decided to. All the derived equations for LSTM are shown below.

$$f_t = \boldsymbol{\sigma}\left(x_t U^f + h_{t-1} W^f\right) \tag{2.11}$$

$$i_t = \boldsymbol{\sigma}\left(x_t U^i + h_{t-1} W^i\right) \tag{2.12}$$

$$C'_t = \tanh\left(x_t U^g + h_{t-1} W^g\right) \tag{2.13}$$

$$C_t = \boldsymbol{\sigma}\left(f_t * C_t + i_t C'_t\right) \tag{2.14}$$

$$o_t = \boldsymbol{\sigma}\left(x_t U^0 + h_{t-1} W^o\right) \tag{2.15}$$

$$h_t = \tanh(C_t) * o_t \tag{2.16}$$

2.4.2 Proposed Learning Model - Deep Mobility

The current cellular network 4G-LTE architecture has a rigid framework and often lacks customization when it comes to offering any tailored business requirements. Service providers have to often rely on third party vendors or equipment manufacturers to provide desirable customer solutions. SONs come into the picture to introduce automatic adaptability in the network like the ANR and MRO functionality does the neighbor list updates and modifications to the hysteresis and TTT values. However, this change is not performed based on the dynamic network or RF changes but based on some service providers' database information from collective HO performances as in HO failures (HOF), call drop rates (CDR), link failure rates (LRF), etc. The current RF conditions might have changed between now and the times these rates were recorded. Some network parameters are modified during maintenance window overnight.

Our model consists of an input layer with number of neurons same as the input

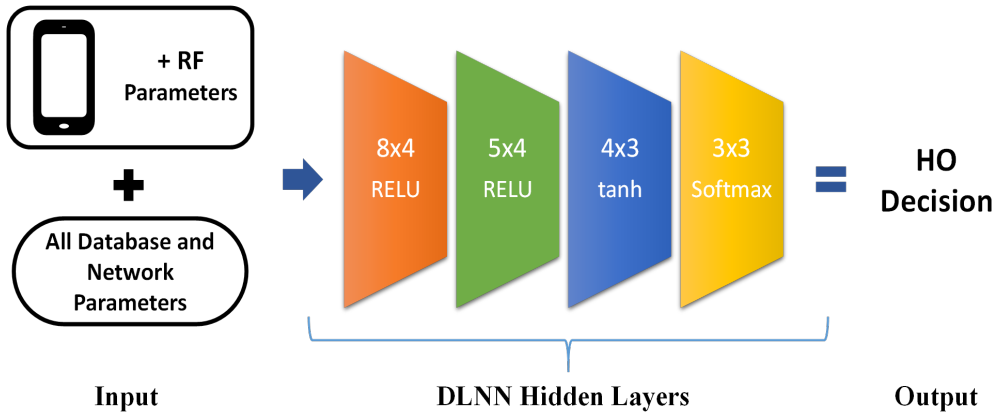

Figure 18: Our Deep-Mobility Neural Network Model

parameters, a couple of hidden layers and an output layer. Each layer is activated using a certain activation function to make parameters non-linear and close to the real world. A 4x3 means a total of 12 neurons in that particular layer. Output layer is applied a simple linear regression function. About 30% dataset was used for validation. Even though the major factor for handovers is movement, there is not one single factor in a real-world scenario which triggers a handover. Even the slightest change in the location in a dense cellular network may cause a user to handover to a neighboring cell. For example, if SNR is the only handoff trigger factor, then maybe we have an acceptable SNR in an area, but the radio resources available are not enough and so we get low throughput in spite of good SNR. If there's an overlapping cell in that region which can provide greater throughput at a slightly lower SNR then the user needs to be handed off to this new cell.

2.4.3 Results

We created a whole dataset of variables from real network deployment for all serving and neighboring base stations which partly contributes to about 65% of our input

dataset used. This UE measurement report information contains the power (RSRP) of serving as well as three to four neighboring cell sites. We make use of multiple third-party mobile applications to retrieve this information; up to four different applications were used to compare and contrast the base stations' data. We also identify these base stations as a 3G (BTS), 4G-LTE (eNodeB) or a 5G base station (gNodeB). Base station identification could be easily done based on the frequency band and the bandwidth reported from these cell sites. Table I shows you some of the basic parameters that are captured by the UE and these Apps for all sites that the device can communicate with. All this was done using multiple 5G and LTE based cell phone devices/UEs.

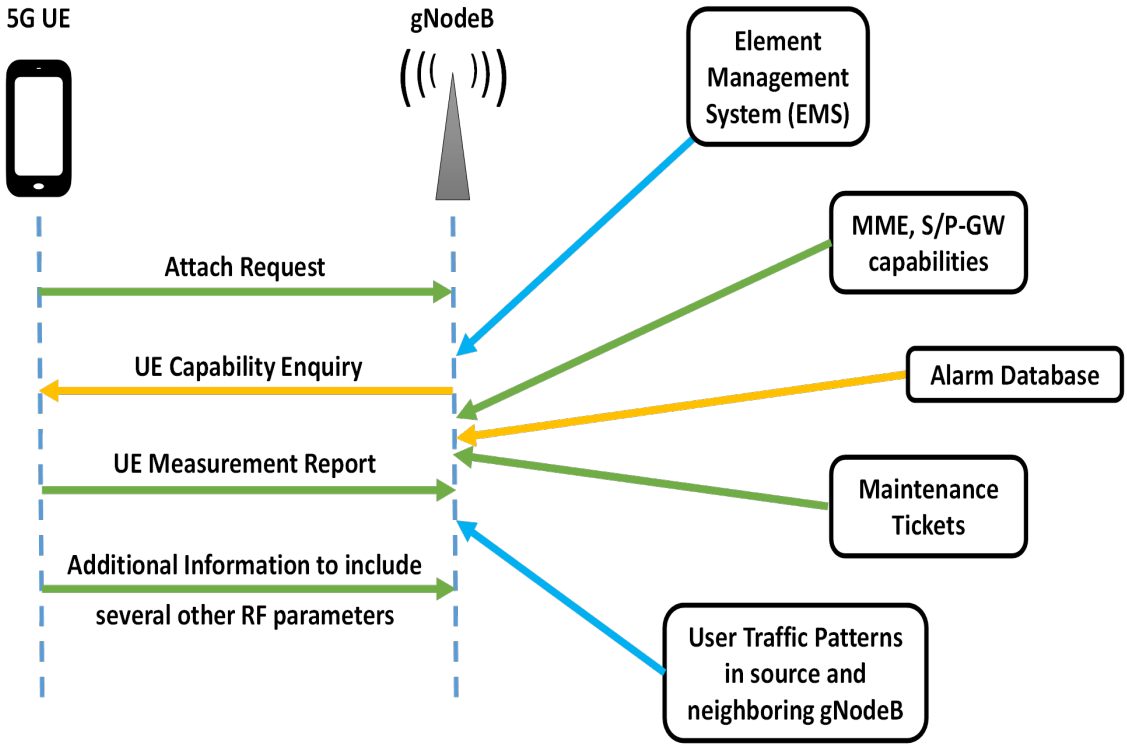

Figure 19: Feature Highlights of UE Measurement Report and Network Centric Parameters

As seen above, applications provide this information with each object having a unique value at every physical location the UE will collect this data from. Most of them are standard and self-explanatory, TAC typically is used to point the BS to a pre-defined pool of Mobile Management Entity (MMEs) and often limited to a geographic area. CID is a unique identification code to determine each sector of each band and is determined as per the service providers' naming scheme. EARFCN is an ETSI industry standard to display the channel numbers instead of raw frequencies in MHz and is bandwidth independent. This is an important parameter in our training process to analyze when was the carrier aggregation (CA) is enabled and under what conditions it was not available. We have other RF parameters which can be captured either by the UE or on the base stations.

In our case, we made some assumptions on these variables for the purpose of training our DLNN. Most of them cannot be accurately determined, unless captured over a long duration of time, like the dynamic RF channel conditions, signal modulation used on the downlink (DL) and the uplink (UL) for each cell BS, scheduling of the Resource

Table 1: Major KPIs Reported by the Mobile Applications

MCC	Mobile Country Code	EARFCN	E-UTRA Absolute RF Channel Number
MNC	Mobile Network Code	RSRP	Reference Signal Received Power
TAC	Tracking Area Code	RSRQ	Reference Signal Received Quality
eNBID	eNodeB Identification	RSSI	Received Signal Strength Indicator
PCI	Physical Cell Identity	CIQ	Channel Quality Indicator
CID	Cell Identity	SID	System Identification
BAND	Cellular Band Used	NID	Network Identification

Blocks (RBs), etc. We have this information captured and shown in Table II for understanding purposes. We would assume some of these neighboring sites would have future maintenance tickets open and are reflected in the database system. We also have a real time monitoring performed by the Element Management System (EMS) which will have present alarms reported from the BS. SONs will be pre-programmed to take action on regular alarms but we don't make that assumption and consider manual intervention will be necessary to fix any alarming issues. Some alarms can be fixed remotely and some require dispatching of a crew to the cell site. Our dataset is not concerned with the alarm fixing process but will only categorize if an existing alarm is serving impacting or not. We flag all serving impacting alarms in our input dataset so our DLNN can differentiate and learn their impacts on users.

We assume a UE-K in our network and in connected mode, our 'Deep-Mobility' model has learned about the movement patterns of this UE-K for the past few weeks and understands that for this particular time, UE-K will be pretty much in a half mile radius of

Table 2: Additional network side attributes in HO decision

EMS	Alarm 1	Allowed
EMS	Alarm 2	Service Impacting
Maintenance Database	Ticket 1	Allowed
Maintenance Database	Ticket 1	Service Impacting
KPI Capturing Tool	CFR	Call Failure Rate
KPI Capturing Tool	CDR	Call Drop Rate
KPI Capturing Tool	HOF	Handover Failure Rate
KPI Capturing Tool	RLF	Radio Link Failure Rate
KPI Capturing Tool	Cell Load	Number of Connected Users

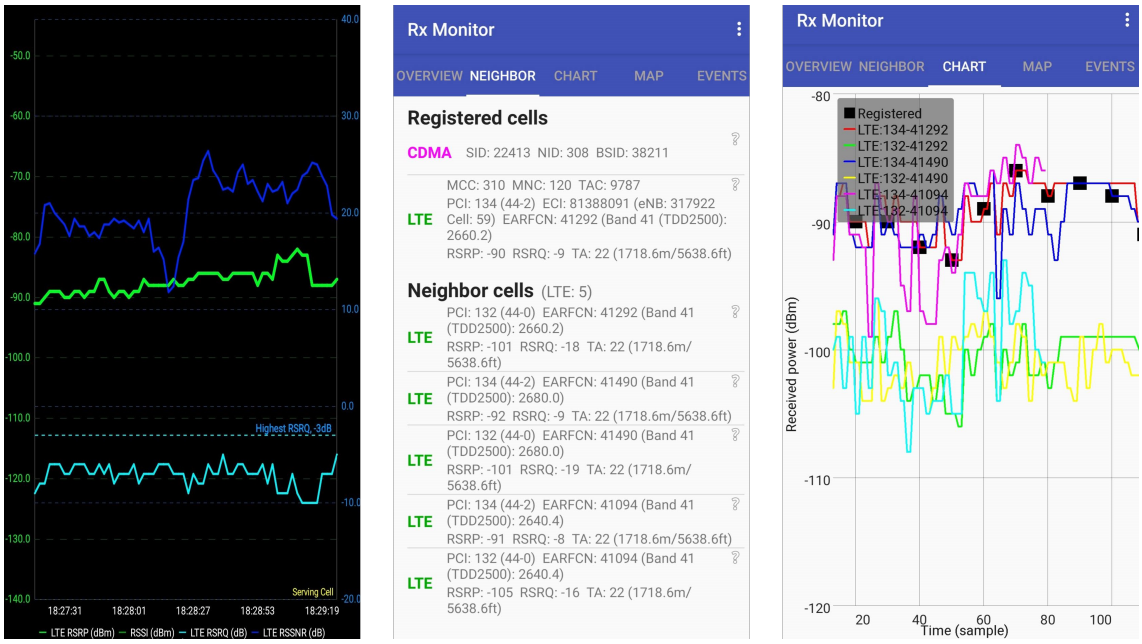

Figure 20: Sample snapshot of Application Reports

the current reported location. Assume that better RSRP and RSRQ conditions are reported for the α (alpha) sector of a target eNB-T than an existing β (beta) sector of the serving eNB-S. Then the network or the UE would initiate a handoff request to eNB-T. What if the EMS has reported a service impacting alarm on that $/beta$ sector of eNB-T which the UE is not aware about, and the network, even when aware, would otherwise not consider this during a HO decision. The UE is better off by not handing over to the eNB-T as it seems to be immobile at this time and location based on the system learnings.

Figures 8 [a, b, c] show glimpses of what we could capture on the mobile applications regarding RSRP, RSRQ and RSSI. It also shows the technology UE is connected to. Our dataset includes the most relevant KPIs from both network and device side, including the type of device connected, QoS Class Identifier (QCI), packet delay budget, maximum

packet loss, time and day of the week, etc. These KPIs are captured from control packets between the UE and network. Since our model will run internally on the network, all this information is readily available. Base on the training and testing we can plot in Fig. 21 and Fig. 22 the accuracy and loss for each of these activities. Accuracy of close to one is a very efficient training. Loss was minimum. Training of 'Deep-Mobility' model involved the entire dataset and validation used 30% random data from the input dataset.

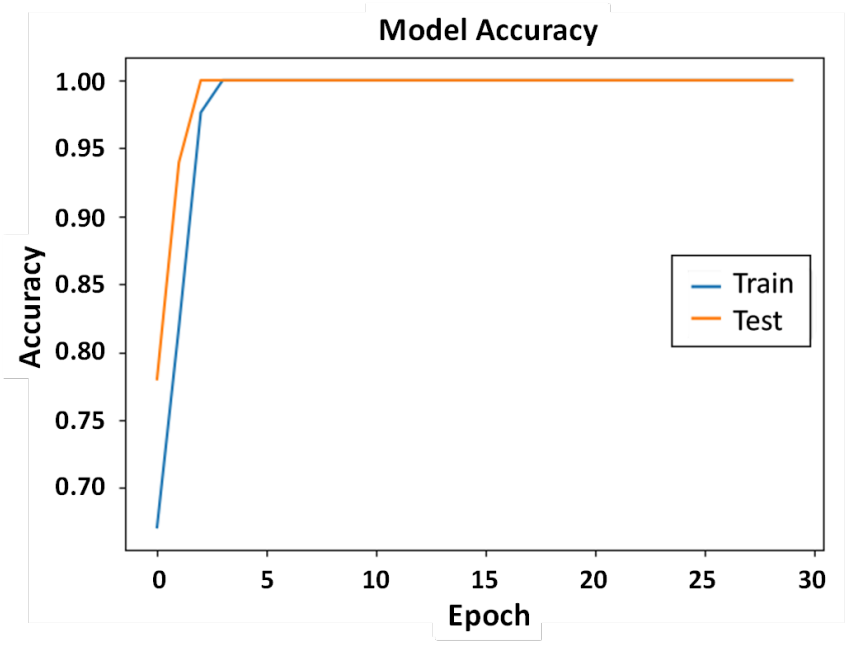

Figure 21: Model Accuracy for Training and Validation

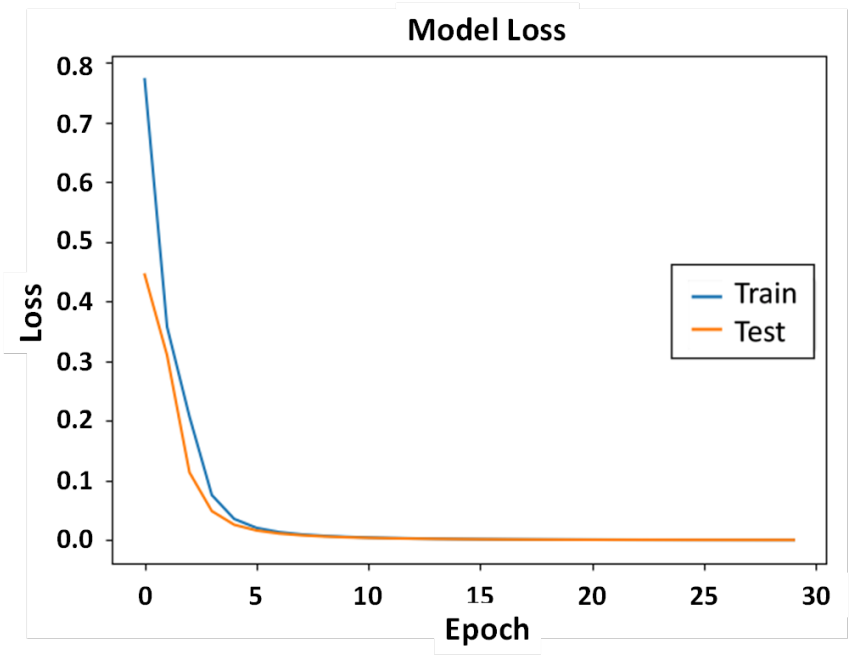

Figure 22: Model Loss for Training and Validation

2.5 Conclusion

The stochastic properties of handover traffic are not simple (i.e., they are not exponential) and have a profound effect on blocking probabilities. The ability to represent a handover connection arrival as a Matrix Exponential will be extremely beneficial in designing efficient handover algorithms. All the network related parameters and mobility dynamics can be studied together in a single model. This analytical model we have proposed is straightforward and flexible yet captures all the detailed information for the parameters causing a cellular handoff. Future work will be to create formulations without using renewal assumptions and to match B matrix formulations to capture real-world traffic, wireless signal propagation, and user movement dynamics. Work that has modeled

some of such situations using Markov chains could be built upon [36]. This will be especially useful for formulating the next generation of 5G handover algorithms that will need to incorporate hyper-network densification, M2M traffic, cloud RAN, and ultra-reliable low latency communication.

The next generation of wireless networks will have multiple cell sites, and HO decisions will be very critical to reliable connections and sufficient QoS. Mobility management and user experience will be of utmost importance for a service provider, and use of Deep Learning to make intelligent network decisions will be a key in shaping the management of future autonomous and self-sufficient network. We designed a novel base model to learn and analyze the UE and the network parameters together and make the most optimal decisions in terms of the user mobility. We accommodate just as many variables as required to accurately predict handovers based on new input parameters fed by the UE or network. Our proposed model is straightforward and flexible, yet captures all the detailed information for the parameters causing a handoff. This will be useful for formulating the next generation of 5G/6G handover algorithms that will need to incorporate hyper-network densification, mmWave, M2M traffic, and ultra-reliable low latency communication.

CHAPTER 3

SURVIVABILITY MODELING IN CELLULAR NETWORKS

3.1 Introduction

This chapter is based on our paper published on survivability modeling of cellular networks, "Survivability and disaster recovery modeling of cellular networks using matrix exponential distributions" by H. Kaja, R. A. Paropkari, C. Beard, and A. Van De Liefvoort, published in IEEE Transactions on Network and Service Management. I want to thank my co-authors for their contribution towards this research.

With the increase in network demand, service providers are focusing on effective ways to increase network capacity and coverage in various ways such as densifying their network, designing effective spectrum reuse policies and aggregating additional spectrum [37]. In order to uphold the QoS metrics of a wireless network, designing the network to ensure survivability and reliability has become critical. A more stringent reliability requirement is also included as one of the key performance indicators (KPIs) in the 5G service requirements as a part of 3GPP Release 15 [38]. An example of such a network is FIRSTNET, a United States nationwide LTE-based public safety network which relies upon highly survivable and prioritized services [39]. This work presents an in-depth evaluation model for survivability of the cellular network in a disaster scenario.

Design of any communication network must include a performance evaluation for a disaster scenario, be it natural or man-made. In a typical disaster scenario, loss in com-

munication is mainly due to base station (BS) failure. A disaster (earthquake, tsunami, hurricane, or man-made attack) could result in a temporary or a long-term infrastructure failure. The theory of disaster propagation shows that calamities propagate into the area with time and grow, thus disrupting more network elements over time. For example, the 2005 hurricane Katrina in Louisiana, USA caused a 15% outage of the network in the first 24 hours [40]. The March 2011 earthquake and tsunami in east Japan damaged 1.9 million fixed-lines and 29 thousand wireless base stations [41]. In-depth disaster analysis is an ongoing research problem and lacks a properly defined approach for all situations, and related cellular network disaster analysis is an important element. Survivability is the resistance of a network to withstand a massive undesired event by maintaining operations or quickly restoring all connections and critical services just as right before the disaster struck [42].

Structured Markov chains capture the dynamics and dependencies of BS failure, recovery activities, and can be presented in many ways. Markov modeling continues to be a widely used tool in survivability analysis, especially when Matrix Exponential representations are used to support arbitrary state-residence distributions instead of simple exponential distributions that are limited in their applicability, particularly for repair processes. This work discusses the computation and analysis of a Markov chain-based ME recovery model for a cellular network to assess network performance during a disaster response. Also, different repair strategies which cater to failed BSs in networks with one or more repair crews has been reviewed. Analysis of such networks could effectively be accomplished through both simulation and analytical models. The ME analytical approach

here is especially helpful compared to simulation because it can find asymptotic behavior that would scale to very large networks that could not be simulated, can implement a wide range of repair distributions and processes, and enables one to make informed design decisions related to numbers of crews, ability to conduct fast repairs, and modification of particular repair process parameters, such as travel time to failed BSs.

This chapter proposes the use of ME distribution in the repair process of a Markov chain so as to capture the multi-dynamic nature of repairs. For example, the first fast repair type can be general quick fixes performed manually and pushed remotely over a network connection. A repair process may or may not include travel to a site by a work crew. And a slower crew-based repair may include travel, diagnosis, obtaining replacement parts, and installation. In both fast and slow repairs, the rate at which each stage of repair is performed is taken into consideration and a state repair matrix ($\mathbf{B}$) is formulated with the corresponding rates. With the use of this $\mathbf{B}$ matrix, an ME distribution model has been constructed, and the characteristics of these repair models are studied in detail. This work considers fast repairs to be associated with a normal failure scenario which includes daily software upgrades, downtime due to routine maintenance checks, minor system errors, etc., whereas slow repairs are associated with more involved scenarios which would require multiple recovery stages and repair crew dispatch to the site locations.

Also, use of an ME distribution allows one to accurately calculate both the transient response and steady state probabilities. For simplicity purposes initially, we assume two types of major repairs, first are those that have a single step of repair (fast repairs). With the development of a cloud based Centralized Self Optimizing Network (CSON),

some or most of these manual interventions are no longer required, and many minor fixes are now automatically compensated for, hence resulting in quite fast repairs. The second type are those that require dispatching a repair crew to the cell site and involve multiple repair steps (slow repairs).

Massive losses of BSs can be the result of a natural or a man-made disaster in a geographical region. Multiple repair crews operate on the recovery of the network after such an undesired event has occurred on a massive scale, unlike in the case of regular failures where only a few repair crews work at a time. This work contributes towards incorporating these types of in-depth analyses on disaster recovery procedures into an ME distribution model and performs a detailed analysis. It then shows how survivable networks are designed with these models.

- Matrix Exponential models that can accurately and analytically represent the failure and repair processes of cellular networks.
- Survivable cellular network design by making design decisions on (1) numbers of available repair crews, (2) balance of fast and slow and repairs, and (3) improving time taken for the steps in the repair process, illustrated by considering improved travel time.
- Transient analysis of network availability after a major event.
- Demonstration of the value of the analytical model to predict performance of very large networks.

The rest of the chapter is organized as follows: Section 3.2 presents related re-

search contributions by various other researchers working in the area of disaster analysis and survivability in cellular networks, Section 3.3 presents our proposed network model and further discusses our key research contributions, Section 3.4 discusses the ME distribution model for disaster recovery in detail along with discussion about various repair models considered. Section 3.5 uses the model for transient analysis, shows how the model extends for very large networks, uses the model for survivability design, and the chapter is concluded with a summary and future work opportunities.

3.2 Related Work

The effects of a disaster like Hurricane Katrina, and the Japan earthquake were studied in depth for interrelationships in telecommunication systems and the power grid [40]. All the damage as well as the restoration was studied by a site survey in [41]. Various studies have been conducted on possible improvements in disaster recovery plans for speedy recovery of telecommunication systems in [43] and [44]. However, a comprehensive statistical analysis of a disaster recovery process is a research topic which is yet to be studied in depth.

Studies have also been performed on network survivability analysis and assessment. Authors in [45] compute the network survivability and propose different models based on the framework for telecommunication network survivability performance. Link and node level survivability was studied in detail for telephone subscriber networks in [46]. Even in the absence of major disasters, wireless network survivability is vital to the expansion of wireless capabilities to support critical infrastructures and applica-

tions. Important examples where highly available wireless networks will be necessary include smart grid, air traffic control, public safety operations (police, fire, ambulance), and remote medicine.

Most related studies have considered exponential distribution models for disaster recovery procedures. A similar approach towards survivability analysis is discussed in [47], using exponential distributions to model failure and recovery processes. Using exponential distributions lacks detail regarding computation of the recovery rate for a step-by-step recovery process and hence limits the depth of knowledge built into the model. Matrix Exponential (ME) models provide scalable complexity to model virtually any situation and probability distribution [34]. Their complexity can be managed depending on the accuracy and metrics of interest. Then the survivability of the network can be engineered effectively.

With known component failure and repair rates, [48] proposes a numerical method based on the Taylor series expansion of the underlying Markov chain stationary distribution associated with the reliability and reward models, to propagate parametric uncertainty to reliability and performability indices of interest. The hierarchical architecture of UMTS networks is modeled in [49] using stochastic models such as Markov chains, semi-Markov processes, reliability block diagrams, and Markov reward models. The models can be tailored to evaluate beyond third generation cellular networks' reliability and survivability attributes as well. A unifying framework is proposed in [50] for state-based models for architecture-based software reliability prediction. The state-based models considered are the ones in which application architecture is represented either as a discrete time Markov

chain (DTMC), or a continuous time Markov chain (CTMC).

In this work, we perform our disaster failure and recovery models using ME distributions and provide observations of a large range of interactions involved in the recovery processes of BSs. Authors in [51] present a survey on the developments and applications of the finite Markov chain imbedding approach in the reliability field consisting of analyses of system reliability, reliability computations, importance analyses of components, failure rate function computations, the birth of new reliability system models, and analyses of new reliability tests. Assuming that an equipment's unobservable degradation state transition follows a Markov chain, authors in [52] design a hidden Markov model to calculate the reliability function, and the mean residual (remaining) life of a piece of equipment, when its degradation state is not directly observable.

Regarding extensions of Markov chains, the authors in [53] work with ME models on arrivals and departures of a multi-server model to computer the first three moments to characterize the performance measures. In [54], the researchers have considered the numerical computation of the stationary distribution for the level dependent quasi-birth-and-death (QBD) process. Research involving detection of the most vulnerable area in a given network where a disaster is more likely to hit in a certain predefined pattern is in [55] and [56]. The statistical placement of BSs in [57] shows the significance of the Poisson point process (PPP) to model the BS and user placement in any given cellular network. The authors also confirm that this technique yields very accurate results when compared with real world deployment [58] [59]. Authors in [60] analyzed disaster-based survivability using a truncated continuous time Markov Chain (CTMC) model. A tran-

sient system analysis of the whole network under a disaster scenario was performed from its fully failed state to when it was restored completely in [61] [62]. Authors in [63] obtain results involving the estimation of the characteristics of functioning of a reliability system modeled by a homogeneous semi-Markov process.

There has been considerable amount of research surrounding different strategies to quicken the process of communication system restoration in an event of a disaster. Some of the ways include cellular service providers deploying temporary BSs such as movable cell sites [64]. A recent development in the field of disaster recovery has been the use of unmanned aerial vehicles (UAVs) assisting in different aspects of the disaster restoration process [65] and cells on wheels (COWs) which can be rapidly deployed in emergency and disaster struck regions [66] [67] [68]. In case of a complete infrastructure failure where some core network elements are also severely affected, device to device (D2D) communications is proposed to be employed [69] [70].

The Markov chain has been a commonly used tool in modelling the survivability of an infrastructure-based network [71] [72]. Our novel use of ME models for matrix exponential transitions in the Markov chain greatly expands the capability of Markov chains to match real-world survivability scenarios. Then it is readily used for accurate survivable network design. Additional background and related work on ME models is provided in the Appendix.

Table 3: Mathematical Notations

$f(t) = \mathbf{p}e^{-\mathbf{B}t}\mathbf{B}\mathbf{e}'$	Generic matrix exponential probability density function
$\mathbf{p}, \mathbf{B}, \mathbf{e}$	For an ME distribution the initiation vector, the rate matrix of progress, and the summing vector
m	The dimension of a $\mathbf{B}$-matrix of a generic ME distribution
$\mathbf{V}$	The inverse $\mathbf{B}$. The first moment of the distribution is $\mathbf{pVe}'$.
$\mathbf{p}_{\#}, \mathbf{B}_{\#}, \mathbf{e}_{\#}$	A particular distribution is identified. For instance, $\mathbf{p}_{\text{B5}}, \mathbf{B}_{\text{B5}}, \mathbf{e}_{\text{B5}}$ is the base model with dimension 5
α_1, α_2	The generic probabilities that a repair is either fast or slow in the base model
$\mu_1, \mu_{2a} \ldots \mu_{2e}$	Rates of various stages in the base model
N	The number of base stations
ν	Failure rate of an active BS
r	The number of repair crews
$\pi_0 \ldots \pi_N$	Stationary probabilities for 0 through N active BSs
$\pi_0(t) \ldots \pi_N(t)$	Transient probabilities for 0 through N active BSs at time t
$\mathbf{Q}_r$	Infinitesimal rate matrix for the Markov Chain for a system with r repair crews.
$\otimes$	Kronecker product, needed to represent multiple active non-exponential repairs
$\mathbf{I}_i$, $i = 1 \ldots r$	Identity matrix for i simultaneous and active repairs
$\mathbf{C}_i$, $i = 1 \ldots r$	Progress rate matrix for i simultaneous and active repairs
$\mathbf{D}_i$, $i = 1 \ldots r$	Rate matrix for the first completion of i simultaneous and active repairs
$\mathbf{E}_r$	Rate matrix for the first completion of r simultaneous and active repairs, with the start of another new repair
$\mathbf{U}_1 \ldots \mathbf{U}_N$	Generic component of a steady state solution

3.3 Model Description

3.3.1 Overall Network Model

Fig. 23 represents a cellular network with BSs placed at different locations providing connectivity to different types of outdoor cellular users, transportation vehicles, household devices or appliances, etc.

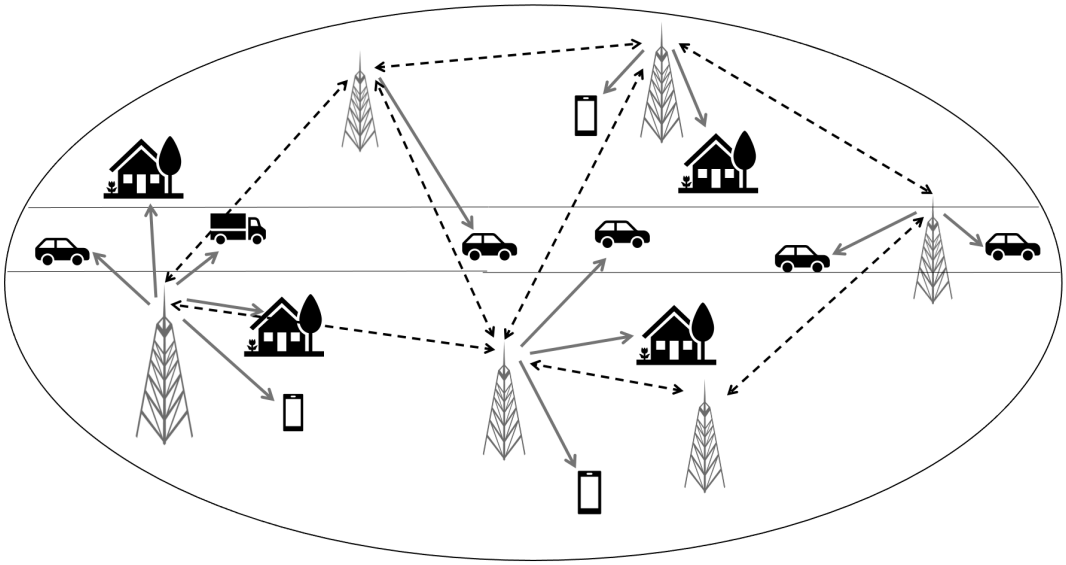

Figure 23: Illustration of the Heterogeneous Network

Table 3 provides a list and description of parameters. For the cellular network being discussed in this chapter, assume the total number of cellular BSs to be N. After a disaster, n denotes the number of active cellular BSs in the network that survived the disaster. The disaster and recovery model uses an ME-based Markov chain where BSs fail according to exponential distributions with failure rates of ν. It is common and considered generally accurate in survivability work to model failures using exponential

distributions [73]. The memoryless property of the exponential distribution approximates the approximately memoryless property of network failures.

The model described in this work can be used to quantify the repair process for any base station network. The repair processes of the base stations are represented by the Matrix Exponential (ME) distributions instead of pure exponential distributions since repair processes frequently depend on a series of steps being carried out. The ME model provides the ability to model a general distribution rather than a singular exponential repair rate. This general distribution can include multiple types of repairs (fast, slow, etc.) and repair processes with multiple stages. The single-parameter memory-less exponential distribution is not sufficient to model general non-exponential distributions or capture the multiple steps in a repair process. Instead, all of these cases can very well be analyzed and studied with ME distributions, the details and derivation of which are given in the following sections. The Markov chain model with exponential failure rates and Matrix Exponential repairs is shown in Fig. 24, where state i represents the number of BS operating correctly, (and thus $N - i$ have failed), with a single repair crew non-exponential process.

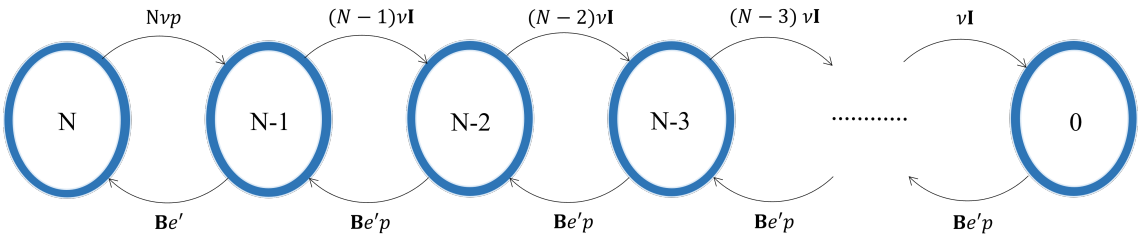

Figure 24: Markov Chain of a System with N Base Stations and One Non-Exponential Repair Crew. The left-most state represents all N Base Stations operating correctly

3.3.2 Survivability, reliability and availability in cellular networks

Survivability in wireless networks is the ability of the network to recover into a basic operational stage after the occurrence of a natural or a man-made disturbance [74] [75], for example enough base stations to cover some percentage of a geographic area. Whereas reliability is defined as ability of a network to be operational for a specified amount of time [75]. Availability has similar definitions but in availability the operation of the network is ensured by adding external or failover resources if required. To distinguish between availability and survivability, availability of a network considers the routine downtime of the hardware components while calculating the total time, whereas, survivability is associated more with a catastrophic disaster and recovery process of the network [76]. All three definitions mentioned above are interlinked and account toward overall dependability of the communications network. To calculate the survivability of the network, both quantitative and qualitative analysis is required. There has been significant work done towards survivability analysis of cellular networks. The authors in [77] discuss a similar approach towards survivability analysis and use the exponential distributions which eliminates the capture of the dynamic nature of the network. It is important to develop an effective network survivability model to capture the dynamic behavior of the network. This work approaches this problem by using ME models with variations in repair processes, number of network repair crews, etc.

This work assumes that base stations fail independently and their repair brings them back to being able to provide service to customers. We do not consider correlated backhaul failures that would also need to be repaired before service is restored to

customers. This independence assumption is most applicable to wind-oriented failures (storms, tornadoes, hurricanes) where buried backhaul would not be so much affected. It may not be as applicable for floods, earthquakes, or for small cells that rely on backhaul connectivity through macro cells. Future work to modify the ME Markov chains presented here, however, could account for such correlated failures. This model has both general benefits to many application areas and specific benefits to the recovery process of cellular networks. On one hand, the model here is general and could also be very useful for other types of facilities failures (power grid substations or power lines, weather or aircraft radar sites, etc.) that could have independent failures, fast or slow repairs, and multiple repair crews. On the other hand, this work is particularly helpful regarding cellular base stations, because they are likely to have independent failures, and fast repairs are more possible through remote fast repair capabilities.

3.4 Performance Model

3.4.1 Matrix Exponential Distributions

Matrix Exponential (ME) distributions are probability distributions whose density functions is represented by

$$f(t) = \mathbf{p}e^{-\mathbf{B}t}\mathbf{B}\mathbf{e}' \quad \text{for } t \geq 0 \tag{3.1}$$

where $\mathbf{p}$ is an m-dimensional row vector (the starting vector), an $m \times m$ matrix $\mathbf{B}$ (the progress rate generator), and an m-dimensional column vector $\mathbf{e}'$ (the summing or closing vector), which in this work consists of all 1's. A number of well known distributions have ME representations, like the exponential, Erlangian, hyper-, and hypo exponentials,

phase types, Coxians, generalized hyper Erlangians, and also their mixtures and convolutions. An ME distribution carries with it the simplicity of capturing Markov chain analysis where instead of a single exponential parameter the transition rates are modeled by matrices. This work primarily exploits ME capabilities to model steps in a repair process, but many distributions to represent a repair process can be used as well. Additional details are discussed in the appendix.

ME distributions can capture all the factors in the recovery of a single failed BS. Various reasons can cause a BS to fail, such as bad weather conditions/natural disasters, network failures, physical link failures, network congestion and overload, configuration issues, etc. [78]. The repair process, similarly, must be flexible enough to capture various possible scenarios. We classify them as a mixture of fast and slow repairs. Fast repairs are related to short time cell outages (STCOs) which are hidden outages. The repairs refer to fast software fixes which can be done remotely. Slow repairs, on the other hand, are related to long term cell outages (LTCOs) which consume more time for the repair process. Slow repairs likely require a crew to be deployed on site to carry out the repair. Fig. 25 illustrates a typical repair process. With probability α_1 it is fast and lasts an ex-

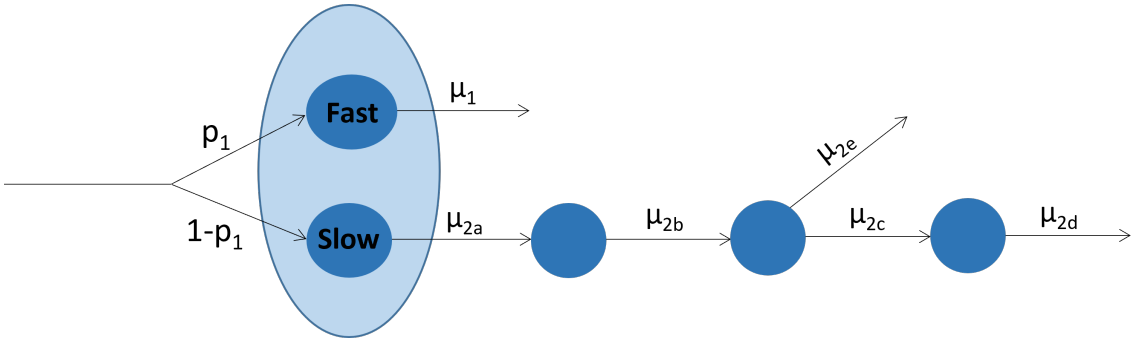

Figure 25: Matrix Exponential Repair Scenario with a Fast Branch and a Slow Branch

ponential time with rate μ_1. Otherwise, it is slow with probability α_2 and consists of a sequence of exponential steps with exponential rates μ_{2a} through μ_{2e}, each step representing sub-tasks, such as travel to a site, diagnosing the cause of failure, and deployment and testing of corrective measures to restore the normal operation of the BS. If any hardware component such as a channel card, router, power distribution unit (PDU) has failed, a replacement part is installed with rate μ_{2e} or, if the spare component needs to be delivered or ordered, the delivery will complete with rate μ_{2c} and installation with rate μ_{2d}. This repair distribution is referred to as the base model with $\mathbf{p}_{\mathrm{B}}$ vector and the $\mathbf{B}_{\mathrm{B}}$ matrix for Fig. 25 which are:

$$\mathbf{p}_{\mathrm{B}} = \begin{pmatrix} \alpha_1 & \alpha_2 & 0 & 0 & 0 \end{pmatrix}$$
$$\mathbf{B}_{\mathrm{B}} = \begin{pmatrix} \mu_1 & 0 & 0 & 0 & 0 \\ 0 & \mu_{2a} & -\mu_{2a} & 0 & 0 \\ 0 & 0 & \mu_{2b} & -\mu_{2b} & 0 \\ 0 & 0 & 0 & (\mu_{2c}+\mu_{2e}) & -\mu_{2c} \\ 0 & 0 & 0 & 0 & \mu_{2d} \end{pmatrix} \tag{3.2}$$

Values for this base model were chosen for illustration purposes only. A real-world study would be needed for practical implementation. These values are $\alpha_1 = 0.60$, $\alpha_2 = 040$, $\mu_1 = 4$, μ_{2a}, $\mu_{2b} = 5$, $\mu_{2c} = 1$, $\mu_{2d} = 1$, and $\mu_{2e} = 0.51$. The mean of this distribution is 0.5633 hour (33.8 minutes) and the squared coefficient of variation (SCV) is $C^2 = 2.00$. The dimension of the underlying matrices is 5, which results in larger matrices in the network model, see below. Therefore, when needed, we use other distributions with a smaller dimension, yet preserve the 60-40 split, and have the same mean and same SCV. We show in Section 3.5 how the results are comparable for these smaller dimension matrices.

Additionally, we study the performance of other distributions, where the mean is again kept at 33.8 minutes, where the split is still 60-40 and where the SCV is $C^2 = 4$ and $C^2 = 8$, see the appendix.

3.4.2 Network Model Description

In the Markov chain representation of a cellular network, the number of active BSs is represented as an oval as shown in Fig. 24. The system is known as an M/ME/r/N/N finite population model, where the repair times are identically and independently distributed with a matrix exponential distribution. It is also referred to as the machine repair model or as the machine interference model, and is also a special case of a Quasi-Birth-Death and similarly structured Markov Chains. Let N be the maximum number of active BSs in the network. The first (left most) oval represents the state that all N BSs are operational and active with 0 broken down, the second oval represents the state that $N-1$ are active and 1 is broken and being repaired, and so on. The zero state denotes that none are active and all broken down.

Consider a disaster in a network that potentially fails all BSs in an area, bringing the active BS count to zero, the right-most oval in the diagram. In addition to the disaster, we assume that after its repair, each BS can fail again with an exponential rate of ν. After the disaster and during the repair, let matrix $\mathbf{B}$ represent a generic ME recovery processes of a single BS repair in the network. The vector $\mathbf{p}$ includes the probability of starting a fast or slow repair. The repair model presented here can have several operating repair crews, where all repair crews are independent and identical to each other. The model in Fig. 24 represents a single repair crew model for service recovery while the model in

Fig. 26 represents a recovery performed by two repair crews in parallel.

The label for each arrow represents the rate of going into the next state. The description of the model with one repair crew (Fig. 24) does not require the use of Kronecker products (explained later in this section) to keep track of any other ongoing repair process since there is only one crew in the system. Once a repair is completed, denoted $\mathbf{Be}'$, this crew becomes available and can start to repair another BS, if needed. This applies to a scenario for everyday repairs. The balance equations for the single repair crew system, Fig. 24, are given in equations (3.3)

$$\begin{aligned} \boldsymbol{\pi}_{N(N\nu)} &= \boldsymbol{\pi}_{N-1}\mathbf{Be}' \\ \boldsymbol{\pi}_{N-1}\left((N-1)\nu\mathbf{I}+\mathbf{B}\right) &= \boldsymbol{\pi}_{N}(N\nu)\,\mathbf{p} + \boldsymbol{\pi}_{\mathrm{N-2}}\mathbf{Be}'\mathbf{p} \\ \boldsymbol{\pi}_{j}\left(j\nu\mathbf{I}+\mathbf{B}\right) &= \boldsymbol{\pi}_{N-1}(j\nu) + \boldsymbol{\pi}_{N-3}\mathbf{Be}'\mathbf{p} \\ &\qquad\qquad \text{for } j = 1\ldots N-2 \\ \boldsymbol{\pi}_{0}\mathbf{B} &= \boldsymbol{\pi}_{1}\nu, \end{aligned} \tag{3.3}$$

where $\boldsymbol{\pi}_i$, $i = 0\ldots N$ are the stationary probabilities. Both the stationary behavior of the system and the transient behavior of some subsystems are computed. The driving equation for the entire system is $\pi(t) = \pi(0)\mathrm{e}^{\mathbf{Q}_1 t}$, where the matrix $\mathbf{Q}_1$ is the system-rate matrix readily identified from the balance equations in equations (3.3). It is a block tridiagonal matrix and the zero-elements are all in blocks of appropriate dimensions. This matrix is shown using the following format.

$$\begin{bmatrix} \bullet & \diamond & 0 \\ \star & \bullet & \diamond \\ 0 & \star & \bullet \end{bmatrix} \text{ as } \begin{array}{c} i \\ 1 \\ 2 \\ 3 \end{array} \begin{array}{c} \text{sub-diag} \\ \\ \star \\ \star \end{array} \begin{array}{c} \text{diag} \\ \bullet \\ \bullet \\ \bullet \end{array} \begin{array}{c} \text{super-diag} \\ \diamond \\ \diamond \\ \\ \end{array} \tag{3.4}$$

The matrix $\mathbf{Q}_1$ is now

$$\mathbf{Q}_1 = \begin{array}{c} i \\ N \\ N-1 \\ \\ j \\ \\ \\ 0 \end{array} \begin{array}{c} \text{sub-diag} \\ \\ \mathbf{B}\mathbf{e}' \\ \\ \mathbf{B}\mathbf{e}'\mathbf{p} \\ \\ \\ \mathbf{B}\mathbf{e}'\mathbf{p} \end{array} \begin{bmatrix} \text{diagonal} & \text{super-diag} \\ -N\nu & N\nu\mathbf{I} \\ -(N-1)\nu\mathbf{I} - \mathbf{B} & (N-1)\nu\mathbf{I} \\ -j\nu\mathbf{I} - \mathbf{B} & j\nu\mathbf{I} \\ \text{for j=1...N-2} & \\ -\mathbf{B} & \end{bmatrix}, \tag{3.5}$$

Consider a small BS network with N=5 BS where the repair time is given by an ME-distribution $\mathbf{B}$ matrix of dimension m=5. The above $\mathbf{Q}_1$ matrix is then of order 26, with 5 partitions of size 5 (when the single repair process is active), and one partition of size 1 (no active repair).

3.4.3 Multiple non-exponential repair crews

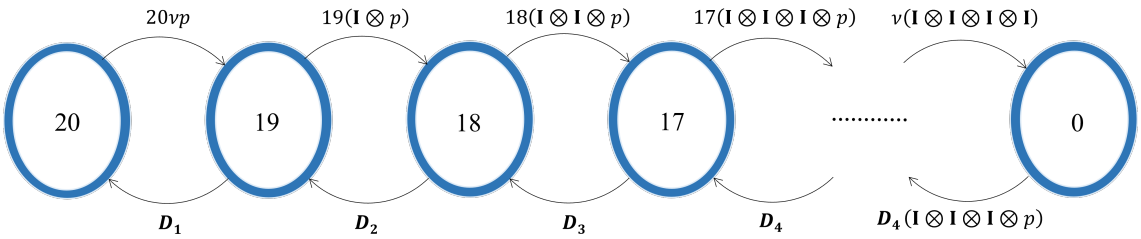

Figure 26: Markov Chain of a System with 20 Base Stations and Four Non-Exponential Repair Crews

When there are two (or more) repair crews available, they work in parallel and independent of each other; see Fig. 26 for an example with 4 repair crews and 20 base stations. Level transitions in the Markov chain occur whenever a BS fails, or whenever a repair is completed.

3.4.3.1 BS failure

When all all BSs are operational, the first failure occurs with rate $N\nu$ and whenever this occurs one repair crew starts the repair and the system enters state $N-1$ with an enlarged space according to $1 \times m$ vector p. When a second BS fails (which occurs with rate $(N-1)\nu$), another crew starts performing repairs without impacting the status of the ongoing repair. The system enters state $N-2$, and its internal state is expanded to two active non-exponential repairs by $(\mathbf{I} \otimes \mathbf{p})$, a $m \times m^2$ matrix, where the identity matrix reflects that the ongoing repair process is not impacted, and where the symbol $\otimes$ represents the Kronecker product representing the expansion. Whenever a subsequent BS fails with i repairs already ongoing and a repair crew is available, the system moves to the next state, with the internal state space again expanded without impacting the ongoing repairs and starting a new repair, denoted by $(\mathbf{I} \otimes \cdots \otimes \mathbf{I} \otimes \mathbf{p})$; there are i identity matrices, and this matrix is of dimension $m^i \times m^{i+1}$. Whenever a subsequent BS fails with a maximum number of r repairs already ongoing and no additional repair crew available, the BS will wait for a crew to become available. The system moves to the next state without impacting the ongoing repairs, denoted by $(\mathbf{I} \otimes \cdots \otimes \mathbf{I})$ with r identity matrices. We define the square matrix $\mathbf{I}_r = \mathbf{I} \otimes \cdots \otimes \mathbf{I}$ for notational convenience, it is of dimension $m^r \times m^r$.

3.4.3.2 Completed repair

A single repair is completed with rate $\mathbf{B}\mathbf{e}'$ and if there are multiple repairs concurrently active, then the first completion will cause a level transition. For 3 active repair

processes, the rate for the first completion is

$$\mathbf{Be}' \otimes \mathbf{I} \otimes \mathbf{I} + \mathbf{I} \otimes \mathbf{Be}' \otimes \mathbf{I} + \mathbf{I} \otimes \mathbf{I} \otimes \mathbf{Be}'. \tag{3.6}$$

This $m^3 \times m^2$ matrix reduces the state space from three active repairs to two active repairs. Should another BS be awaiting repair, it is started immediately and independently by the now available crew, reflected by $(\mathbf{I} \otimes \mathbf{I} \otimes \mathbf{p})$, an $m^2 \times m^3$ matrix. This leaves the two older process intact and starts a new parallel process. There are again 3 repairs active.

3.4.3.3 Balance Equations

In order to concisely describe the balance equations and the rate matrix, we now assume that there are $r = 4$ repair crews only and introduce matrices $\mathbf{C}_i$, $i = 1, 2, 3, 4$ to describe the rate of progress of i simultaneous and independent repairs,

$$\begin{aligned}
\mathbf{C}_1 &= \mathbf{B} \\
\mathbf{C}_2 &= \mathbf{B} \otimes \mathbf{I} + \mathbf{I} \otimes \mathbf{B} \\
\mathbf{C}_3 &= \mathbf{B} \otimes \mathbf{I}_2 + \mathbf{I} \otimes \mathbf{B} \otimes \mathbf{I} + \mathbf{I}_2 \otimes \mathbf{B} \\
\mathbf{C}_4 &= \mathbf{B} \otimes \mathbf{I}_3 + \mathbf{I} \otimes \mathbf{B} \otimes \mathbf{I}_2 + \mathbf{I}_2 \otimes \mathbf{B} \otimes \mathbf{I} + \mathbf{I}_3 \otimes \mathbf{B}
\end{aligned} \tag{3.7}$$

The dimension of $\mathbf{C}_i$ is $m^i \times m^i$, where i is the number of active repairs and m is the dimension of the repair process. Similarly, the $m \times 1$ vector $\mathbf{D}_1 = \mathbf{Be}'$ is the rate at which a single (and only active) repair process completes. The matrices $\mathbf{D}_i$, $i = 2, 3, 4$ are rectangular of size $m^i \times m^{i-1}$ and describe the rate at which the first of i repairs

completes,

$$\begin{aligned}
\mathbf{D}_1 &= \mathbf{B}\mathbf{e}' \\
\mathbf{D}_2 &= \mathbf{B}\mathbf{e}' \otimes \mathbf{I} + \mathbf{I} \otimes \mathbf{B}\mathbf{e}' \\
\mathbf{D}_3 &= \mathbf{B}\mathbf{e}' \otimes \mathbf{I}_2 + \mathbf{I} \otimes \mathbf{B}\mathbf{e}' \otimes \mathbf{I} + \mathbf{I}_2 \otimes \mathbf{B}\mathbf{e}' \\
\mathbf{D}_4 &= \mathbf{B}\mathbf{e}' \otimes \mathbf{I}_3 + \mathbf{I} \otimes \mathbf{B}\mathbf{e}' \otimes \mathbf{I}_2 + \mathbf{B}\mathbf{e}' \otimes \mathbf{I}_3 + \mathbf{I}_2 \otimes \mathbf{B}\mathbf{e}' \otimes \mathbf{I} + \mathbf{I}_2 \otimes \mathbf{B}\mathbf{e}'
\end{aligned} \tag{3.8}$$

Finally, when a repair completes while another BS is awaiting repair, then the $m^4 \times m^4$ matrix $\mathbf{E}_4 = \mathbf{D}_4(\mathbf{I}_3 \otimes \mathbf{p})$ is the rate at which a repair is completed plus a new repair started. Note, that this also places the latest repair to start at the end, so that the repairs are ordered from 'oldest' to 'youngest', see [53]. The balance equations for a system with N BSs and $r = 4$ repair crew are now readily given by

$$\begin{aligned}
\boldsymbol{\pi}_N \, N\nu &= \boldsymbol{\pi}_{N-1}\mathbf{D}_1 \\
\boldsymbol{\pi}_{N-1}\left\{(N-1)\nu\mathbf{I}_1 + \mathbf{C}_1\right\} &= \boldsymbol{\pi}_N N\nu\mathbf{p} + \boldsymbol{\pi}_{N-2}\mathbf{D}_2 \\
\boldsymbol{\pi}_{N-2}\left\{(N-2)\nu\mathbf{I}_2 + \mathbf{C}_2\right\} &= \boldsymbol{\pi}_{N-1}(N-1)\nu(\mathbf{I}_1 \otimes \mathbf{p}) + \boldsymbol{\pi}_{N-3}\mathbf{D}_3 \\
\boldsymbol{\pi}_{N-3}\left\{(N-3)\nu\mathbf{I}_3 + \mathbf{C}_3\right\} &= \boldsymbol{\pi}_{N-2}(N-2)\nu(\mathbf{I}_2 \otimes \mathbf{p}) + \boldsymbol{\pi}_{N-4}\mathbf{D}_4 \\
\boldsymbol{\pi}_{N-4}\left\{(N-4)\nu\mathbf{I}_4 + \mathbf{C}_4\right\} &= \boldsymbol{\pi}_{N-3}(N-3)\nu(\mathbf{I}_3 \otimes \mathbf{p}) + \boldsymbol{\pi}_{N-5}\mathbf{E}_4 \\
\boldsymbol{\pi}_j\left\{j\nu\mathbf{I}_4 + \mathbf{C}_4\right\} &= \boldsymbol{\pi}_{j+1}\,(j+1)\nu + \boldsymbol{\pi}_{j-1}\mathbf{E}_4 \\
&\qquad\qquad \text{for } j = 1 \ldots N-5 \\
\boldsymbol{\pi}_0\mathbf{C}_4 &= \boldsymbol{\pi}_1\mathbf{E}_4
\end{aligned}$$

The stationary probabilities are, for $j = 1 \ldots N-1$

$$\begin{aligned} \boldsymbol{\pi}_{j-1} &= \boldsymbol{\pi}_j \mathbf{U}_j \\ \boldsymbol{\pi}_{j-1} &= \boldsymbol{\pi}_N \mathbf{U}_N \mathbf{U}_{N-1} \ldots \mathbf{U}_j, \end{aligned} \tag{3.10}$$

where the π_N is a scalar (used for normalization) and the matrices $\mathbf{U}_j$ are recursively defined and calculated by

$$\begin{aligned} \mathbf{U}_1 &= \mathbf{E}_4 \mathbf{C}_4^{-1} \\ \mathbf{U}_j &= j\nu \left\{ (j-1)\nu \mathbf{I}_4 + \mathbf{C}_4 - \mathbf{U}_{j-1}\mathbf{E}_4 \right\}^{-1} \\ &\qquad \text{for } j = 2 \ldots N-4 \\ \mathbf{U}_{N-3} &= {}_{(N-3)}\nu (\mathbf{I}_3 \otimes \mathbf{p}) \times \left\{ {}_{(N-4)}\nu \mathbf{I}_4 + \mathbf{C}_4 - \mathbf{U}_{N-4}\mathbf{E}_4 \right\}^{-1} \\ \mathbf{U}_{N-2} &= {}_{(N-2)}\nu (\mathbf{I}_2 \otimes \mathbf{p}) \times \left\{ {}_{(N-3)}\nu \mathbf{I}_3 + \mathbf{C}_3 - \mathbf{U}_{N-3}\mathbf{D}_4 \right\}^{-1} \\ \mathbf{U}_{N-1} &= {}_{(N-1)}\nu (\mathbf{I}_1 \otimes \mathbf{p}) \times \left\{ {}_{(N-2)}\nu \mathbf{I}_2 + \mathbf{C}_2 - \mathbf{U}_{N-2}\mathbf{D}_3 \right\}^{-1} \\ \mathbf{U}_N &= {}_{N}\nu \mathbf{p} \left\{ {}_{(N-1)}\nu \mathbf{I}_1 + \mathbf{C}_1 - \mathbf{U}_{N-1}\mathbf{D}_2 \right\}^{-1} \end{aligned}$$

The dimension of $\mathbf{U}_i$, $i = 1 \ldots N-4$, is $m^4 \times m^4$, while the dimensions of $\mathbf{U}_{N-3}$, $\mathbf{U}_{N-2}$, $\mathbf{U}_{N-1}$, and $\mathbf{U}_N$ are is $m^3 \times m^4$, $m^2 \times m^3$, $m \times m^2$, and $1 \times m$.

3.4.3.4 Transient Solutions

The transient behavior of some subsystems are to be computed as well, $\pi(t) = \pi(0)\mathbf{e}^{\mathbf{Q}_4 t}$, where the matrix $\mathbf{Q}_4$ is the rate matrix for this system and is readily identified from the balance equations in equations (3.9). It is again a block tri-diagonal matrix where the blocks are of appropriate dimensions; only the three non-zero diagonals are shown for

clarity.

$$
\mathbf{Q}_4 = \begin{array}{c} i \\ N \\ N-1 \\ N-2 \\ N-3 \\ N-4 \\ \\ j \\ \\ \\ 0 \end{array}
\begin{array}{c}
\begin{array}{ccc} \text{sub-diag} & \text{diagonal} & \text{super-diag} \end{array} \\
\left[\begin{array}{ccc}
 & -N\nu & N\nu\mathbf{p} \\
\mathbf{D}_1 & -(N-1)\nu\mathbf{I} - \mathbf{C}_1 & (N-1)\nu(\mathbf{I}_1 \otimes \mathbf{p}) \\
\mathbf{D}_2 & -(N-2)\nu\mathbf{I} - \mathbf{C}_2 & (N-2)\nu(\mathbf{I}_2 \otimes \mathbf{p}) \\
\mathbf{D}_3 & -(N-3)\nu\mathbf{I}_3 - \mathbf{C}_3 & (N-3)\nu(\mathbf{I}_3 \otimes \mathbf{p}) \\
\mathbf{D}_4 & -(N-4)\nu\mathbf{I}_4 - \mathbf{C}_4 & (N-4)\nu\mathbf{I}_4 \\
 & & \\
\mathbf{E}_4 & -j\nu\mathbf{I}_4 - \mathbf{C}_4 & j\nu\mathbf{I}_4 \\
 & \text{for j=1...N-5} & \\
 & & \\
\mathbf{E}_4 & -\mathbf{C}_4 &
\end{array}\right]
\end{array}
\tag{3.11}
$$

This $\mathbf{Q}_4$ matrix in (3.11) is the basis for the transient performance metrics. The order of this matrix is $1+m+m^2+m^3\cdots+m^{r-1}+m^r\times(N-r)m^r = \frac{m^r-1}{m-1}+(N-r)m^r$ and increases with the addition of each repair crew, with BS, or with the dimension of the matrix exponential distribution of the repair process. For instance, for m=3, r=4 and N=10 the dimension of $\mathbf{Q}_4$ is 526. The next section discusses our experiments and results with varying number of repair crews and repair distributions.

3.5 Results

This section discusses the results of using the models for survivability design. Models include the number of repair crews in the network, the characterization of the repairs (fast or slow), the balance of fast and slow repairs, individual parameters to describe parts of a repair process, etc. In all the figures below, the initial state is the state of disaster where there are 0 BSs working and as times progresses BSs are revived to operating states. In the graphs below, we consider two time based metrics namely, service restoration time and network availability time. The service restoration time is defined as

the time taken to completely restore the BSs to their normal operating conditions. Here, this has been found by observing the transient state probability for State N (all BSs are operational). We have considered reaching up to 90% of the final stabilized probability to be the service restoration threshold in following figures. Service availability time is defined as the time taken for the network to be available after a disaster, which shut-down the network. This is calculated by observing the transient state probability of a certain number of base stations that are active. In figures below, we consider a network to be available only when 40% of the BSs are active. The repair model which is used for most of the graphs is the $\mathbf{B}_{\mathrm{B3}}$ matrix as given in Eq.(23) in the appendix unless mentioned otherwise. The failure rate of the BSs is taken as $\nu = 0.007$ /hr, as in [74] and [45]. The results section is divided into four main parts. Section-A presents the transient and steady state analysis of the network, Section-B shows the scalability of the ME model being considered, Section-C shows how survivability design is accomplished with this ME model, and Section-D provides insight into the ME model by showing network availability and service restoration time comparisons for different sized $\mathbf{B}$ matrices (with same mean and C^2 values).

3.5.1 Transient and steady state analysis

This subsection discusses the transient and steady state analysis for the chapter's ME survivability model for cellular networks. Each curve in Fig. 27 represents the transient state probabilities of each state of a network with 15 BSs. The system starts with all BSs failed (starting probability 1.0 for state 0) as assumed in [47] to assess survivability and service restoration in the context of widespread failures. Each curve is for the proba-

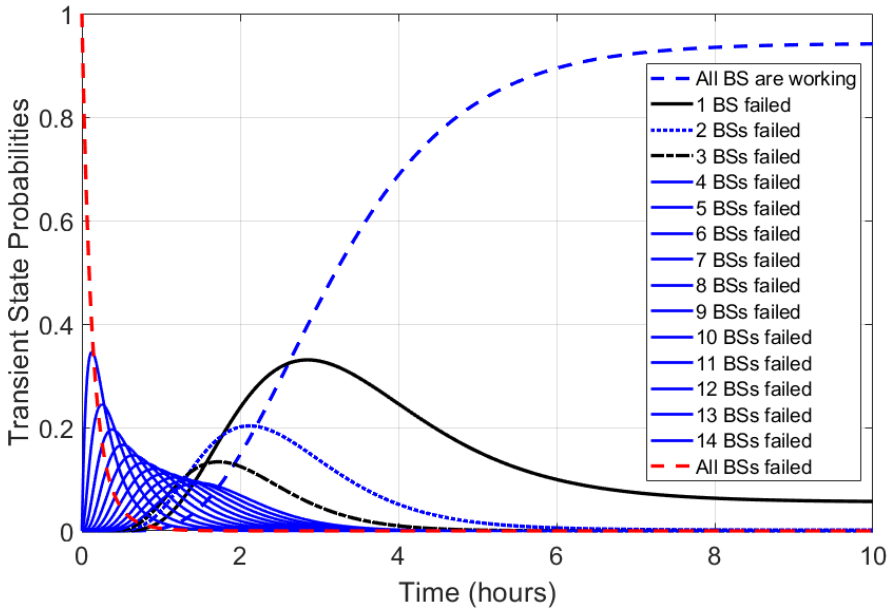

Figure 27: Transient Analysis of a Network with 15 Base Stations and Four Non-Exponential Repair Crews

bility of the number of failed BSs as a function of time. The transient state probability of state N does not stabilize at 1.0, since there is also a significant stationary probability of one base station being inoperable at any particular time. Steady state is then the normal operational state of the network. The probability of going into a fast or slow repair is 0.6 (fast) and 0.4 (slow) respectively, so the majority are fast repairs.

The choices of the parameters of fast and slow repair rates are for illustration purposes only; repair rates would need to be carefully studied for a particular service provider and cellular network. Providers could also use a different process for Fig. 25, resulting in a different set of $\mathbf{B}$ and $\mathbf{p}$ vectors. Survivable network design would also consider how survivability is improved as these parameters vary.

3.5.2 Scalable ME survivability model for very large networks

This section illustrates an ME model which is scalable for networks with higher numbers of BSs. Fig. 28 illustrates the service availability of a network with varying

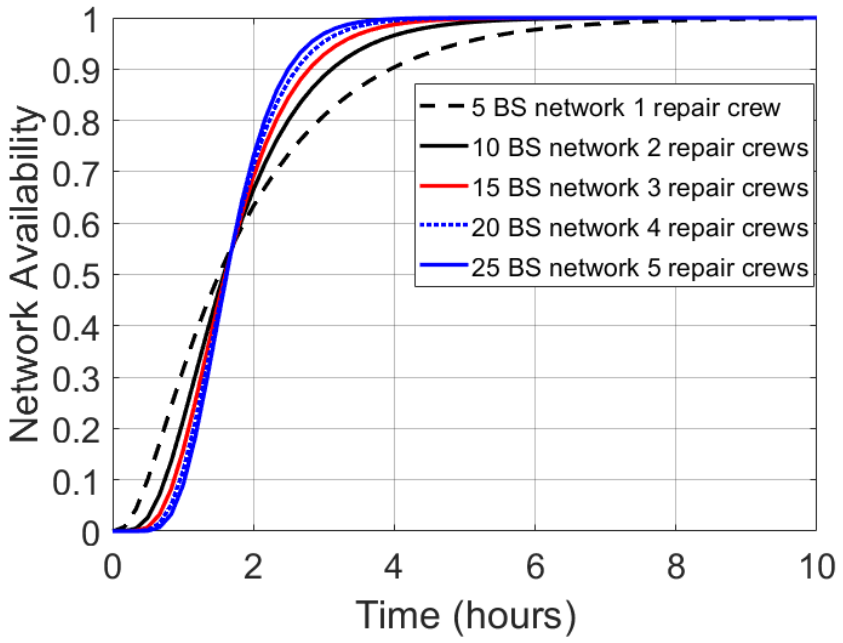

Figure 28: Transient Availability based on the same proportions of BSs and Crews

numbers of BSs and repair crews, but with the same proportion of BSs to repair crews for each curve. For this example, a provider says that a network is available when on average 40% of the base stations are working. Fig. 28 shows the transient response starting with all BSs failed. For each curve, the proportion of BSs to crews is 5:1. As the number of BSs increases, curves become closer, with each larger network causing less marginal change. For example, the transient curve crosses 90% of its final value at approximately 2.5 hours for 20 and 25 BSs, and the curves are very close. This would scale for even

larger networks. For example, for 300 BSs having 60 crews, approximately 2.5 hours to cross that 90% line is also expected. This, in general, provides a procedure where smaller networks of BSs could be evaluated with this ME model using a fixed proportion of BSs to repair crews. Once the smaller networks show very little difference in transient response as they increase in size, the performance of much larger networks is closely approximated. The scalability of this analytical ME model and its calculation approach is extremely useful.

3.5.3 Survivability Design

This work will now highlight three methods that can use this ME model to design highly survivable networks. These involve making design decisions on (1) numbers of available repair crews, (2) balance between fast and slow repairs, and (3) improving time taken for the steps in the repair process, illustrated by considering improved travel time.

3.5.3.1 Impact of numbers of crews on network availability time

Fig. 29 shows the different curves of service restoration for a 15 BS network with 2 to 6 repair crews. This graph illustrates that as the number of repair crews increases, the transient response is faster and restoration times get shorter. Also, the transient response by adding each repair crew stays relatively the same after 5 repair crews. Fig. 30 illustrates the comparison of maximum restoration time and the number of crews in the network. Figs. 29 and 30 demonstrate that initially service restoration time decreases rapidly by increasing the number of repair crews; after 5 repair crews the decrease in the restoration time is not as rapid, only a small improvement is observed.

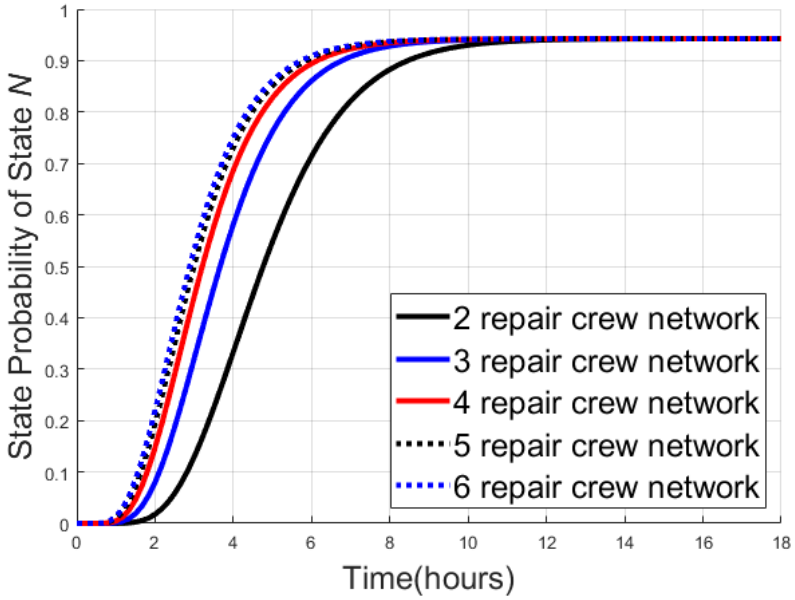

Figure 29: Transient State Probability for State N for a Network with 15 BS and a Varying Number of Non-Exponential Repair Crews

Fig. 31 shows the network availability time for a 15 BS network with varying number of repair crews. Using Fig. 31, we observe that response time and availability improves for a network with higher numbers of repair crews. However, the marginal improvement in availability decreases for each additional crew. Network survivability design would then decide on the balance of the cost of deploying an additional repair crew versus the value of the marginal increase in availability time. In Fig. 31 we observe that the availability time converges to a point with increase in the number of repair crews, similar to what was seen in Fig. 29. This limit is derived below.

Assume the situation where there are N repair crews for an N BS network. This would infer that there is one repair crew per BS and the BSs can be repaired independently. This would mean that each BS goes through an alternating process of being active and

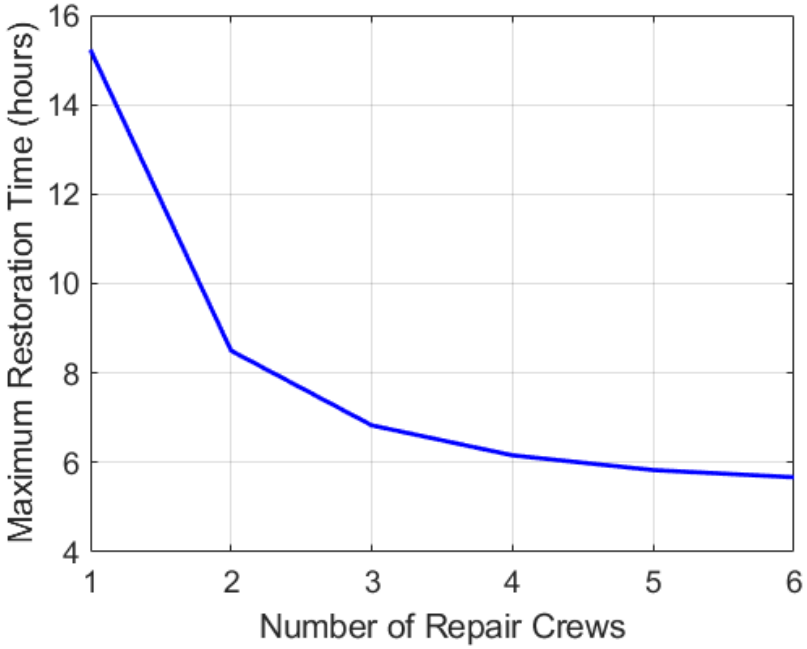

Figure 30: Service Restoration Time for a Network with 15 BS and a Varying Number of Non-Exponential Repair Crews

being repaired. Considering the stationary situation, the average amount of time a BS is operating is given by $1/\nu$ and the average amount of time the BS is being repaired is calculated as $\mathbf{pVe}'$. The probability of each individual BS being operational is given by the expression

$$F = \frac{1/\nu}{1/\nu + \mathbf{pB^{-1}e}'} \tag{3.12}$$

Since BSs are independent, the total number of BSs that are operational is Binomially distributed with F as the probability.

$$\Pr[k \text{ BS's operational}] = \binom{N}{k} F^k (1-F)^{N-k} \tag{3.13}$$

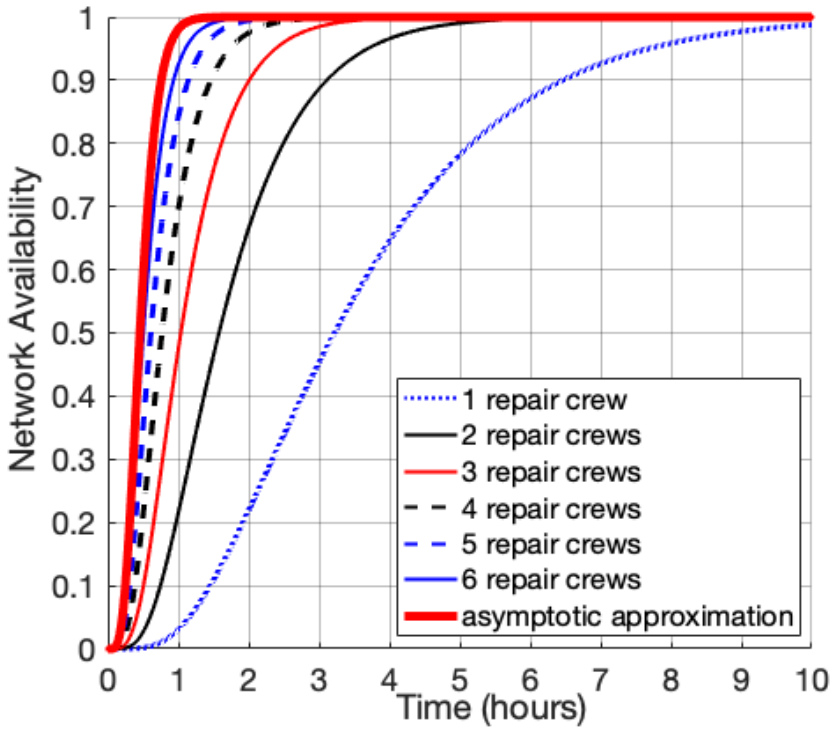

Figure 31: Transient State Probability for a Network with 15 BS and with a Varying Number of Repair Crews

This result also extends to the transient situation as follows:

$$\Pr[k \text{ BS's up at time } t] = \binom{N}{k} (F(t))^k (1 - F(t))^{N-k} \tag{3.14}$$

The $F(t)$ is found by using a stochastic process with infinitesimal rate matrix $\mathbf{Q}_h$ which is defined as

$$\mathbf{Q}_h = \begin{bmatrix} -\nu & \nu\mathbf{p} \\ \mathbf{Be}' & -\mathbf{B} \end{bmatrix} \tag{3.15}$$

If a BS just starts a repair process at time $t = 0$, then at time t the probability that this BS is operational or in repair states is given by the following vector

$$\begin{bmatrix} 0 & \mathbf{p} \end{bmatrix} exp(\mathbf{Q}_h t) \tag{3.16}$$

The probability of a BS being operational at time t is given by post-multiplying a column vector as shown below when the $\mathbf{B}$ matrix is of size 3×3

$$[0\ \mathbf{p}] \exp\left(\mathbf{Q}_h t\right) [1\ 0\ 0\ 0]^T \tag{3.17}$$

The above equations provide the asymptotic curve that is seen in Fig. 31. It is hard to distinguish the difference between the curve where $r = 6$ and the asymptotic curve. So, for results beyond $r = 6$ we can use asymptotic results and we can see there is no value in deploying more than 6 crews.

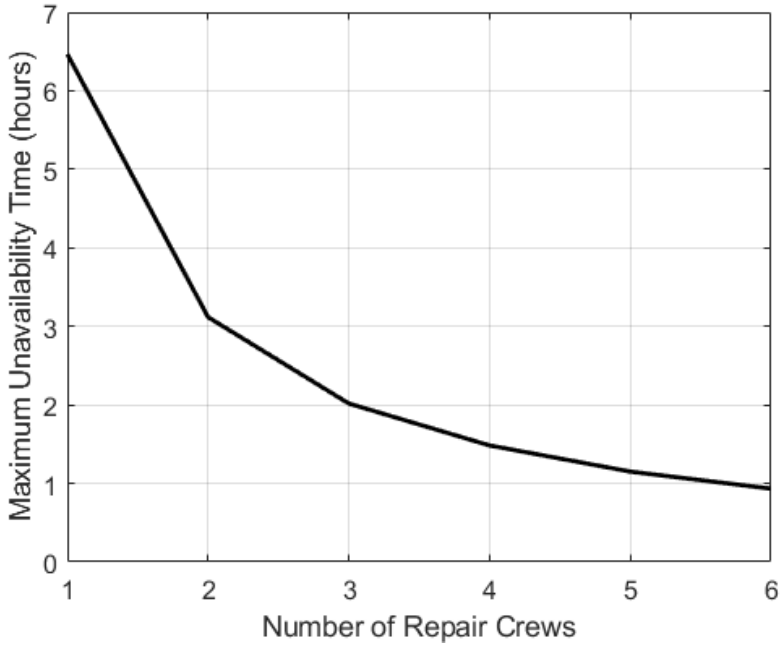

Figure 32: Transient Unavailability for a Network with 15 BS with a Varying Number of Repair Crews

Fig. 32 represents an elaborated illustration of the effect of the variation of the number of repair crews. In this case, transient time is illustrated, as defined as the time

to reach 90% of the final value from Fig. 31. As seen from the plot, initially the transient time decreases rapidly until 4 or 5 repair crews and after that there is a slow decrease in the transient time with increase in repair crews. This graph is especially important for larger networks to determine the required number of repair crews which will reduce the transient time significantly.

3.5.3.2 Balance Between Fast and Slow Repair Processes

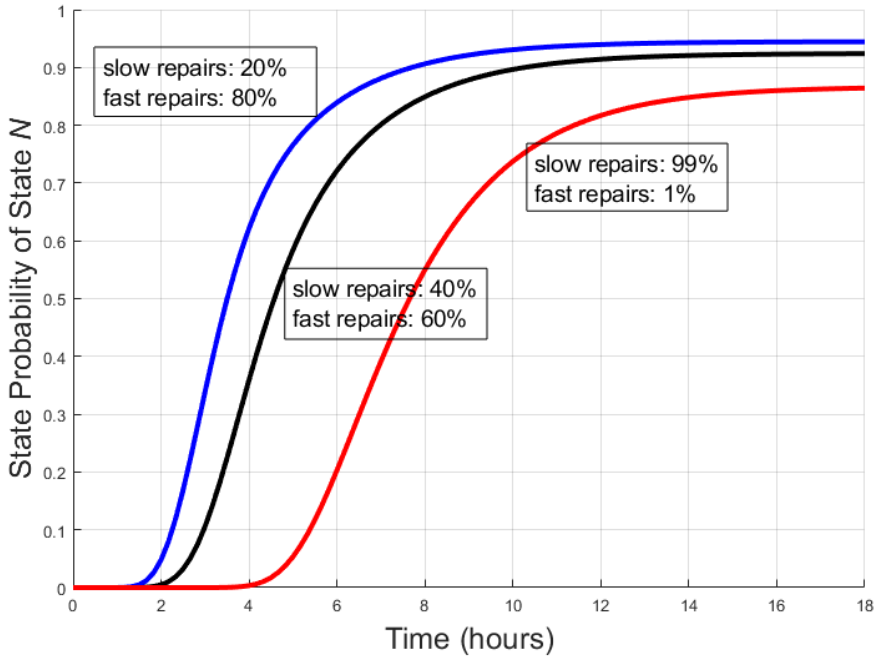

Figure 33: Balance Between the Fast and Slow Repairs for a 20 BS Network with 3 Repair Crews

Now let us consider the benefit of survivable network design to create more opportunities for fast repairs. Fig. 33 illustrates the effects of different ratios of fast and slow repairs being taken in the vector $\mathbf{p}$. In a situation needing 99% of slow repairs, the net-

work restoration time is the longest and the restoration time decreases as the ratio of fast repairs increases as compared to slow repairs. Even though previous figures have started from all BSs failed, still there were a significant proportion of fast repairs. Fig. 33 also shows the full time dynamics that are involved with different balances of fast and slow repairs.

3.5.3.3 Impact of Travel Time Variations

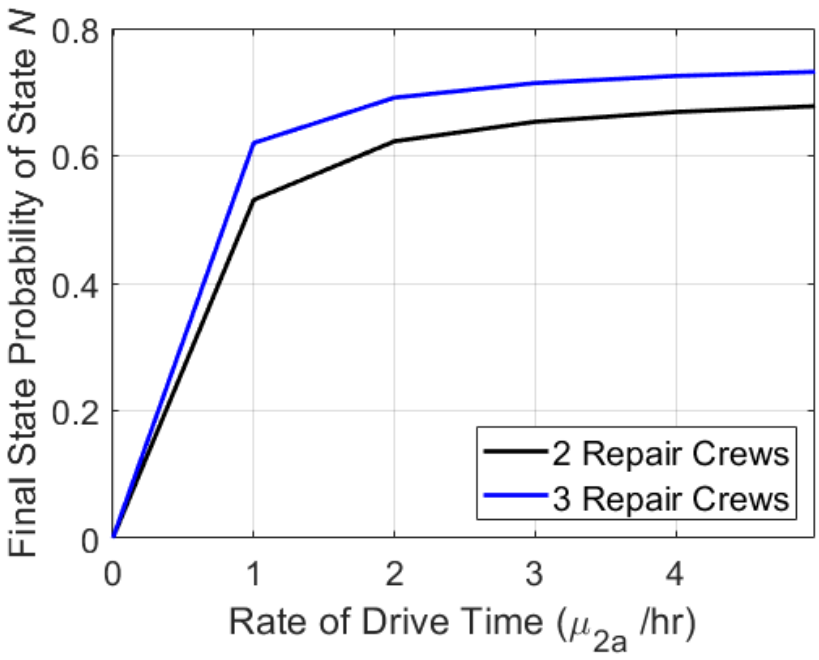

Figure 34: Transient State Probability of Variation in Drive Time Rates for a Network with Two or Three Repair Crews.

Fig. 34 demonstrates how these ME models can quickly show in detail how variations of particular parts of repair processes affect the overall survivability of the network. For example, travel time might be improved by having more service centers where vehi-

cles are located to be ready to respond to a failure. This can be examined in the $\mathbf{B}_{\mathrm{B5}}$ matrix for the travel time parameter μ_{2a}. Fig. 34 shows the impact on final state probability for State N from variation in travel time rates from 0 to 5 per hour. From the graph, it can be inferred that once $\mu_{2a} \geq 2$ per hour (average drive time $1/\mu_{2a} \leq 30$ min.), no further significant benefit is gained. And this is true regardless of the number of repair crews. By using these ME models, a survivable cellular network can be readily designed.

3.5.4 Service Restoration and Service Availability with Different Repair Processes

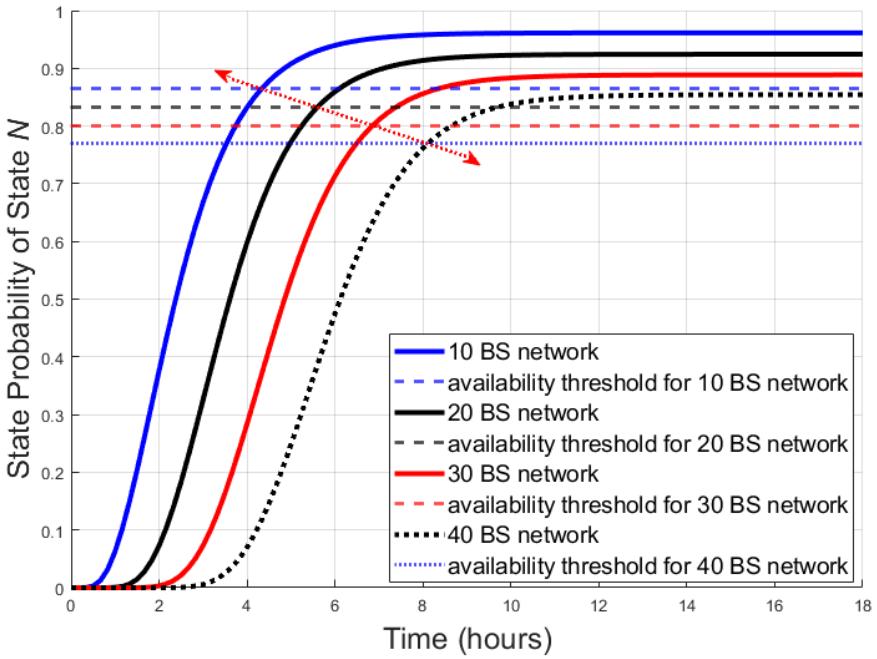

Figure 35: Transient State Probability of State N for a Network with 10, 20 and 30 BSs

In this section we provide a comparison of service restoration and network availability times by varying the network size, using different repair models and varying the SCV values.

The graph in Fig. 35 shows the service restoration times for network models with 10, 20, 30 and 40 BSs and 5 repair crews in each network. As observed, the higher the number of BSs the longer is the restoration time. With higher numbers of base stations, the probability of state N does not stabilize near 1.0, since the probabilities of 1 or 2 base stations failed i.e. states $N-1$ and $N-2$ are significant, as seen in Fig. 27. We also make the interesting observation that the threshold point for each curve aligns on a straight line as shown in the figure. This allows us to estimate the threshold point for larger networks also.

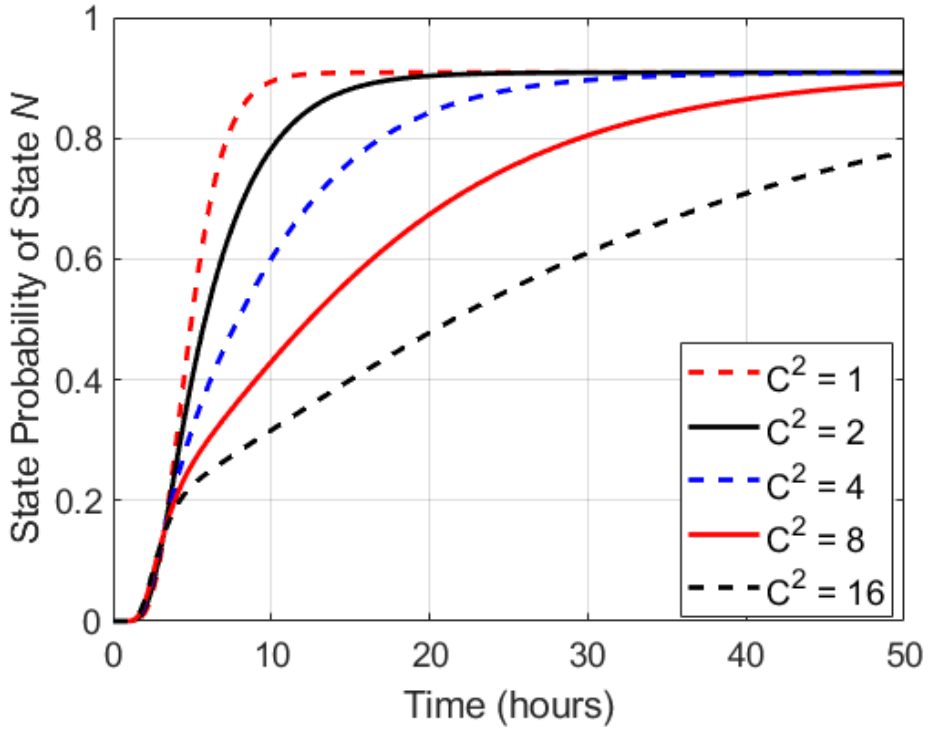

Figure 36: Transient State Probability for a Network of 20 BS with Four Repair Crews and Varying SCV values

Fig. 37 shows a comparison of network availability for a 15 BS network with

varying repair crews and repair models which are two-step slow repairs with a $\mathbf{B}_{B3}$ matrix, three-step slow repairs with the $\mathbf{B}_{B4}$ matrix, and multiple-step slow repairs with the $\mathbf{B}_{B5}$ matrix as indicated in the Appendix. The numerical values of the matrices considered in this graph are provided in the Appendix section and are again compared where availability is considered to defined as having 40% of BSs being available. For the graph, consider a network provider who analyzes network availability using one of the three repair configurations. According to the numbers chosen for this graph, all repair models have same SCV value (which is 2). Also, curves cross as time progresses and converge into a single point at the top end. However, we observe a very small variation in the network availability time in the initial stages while progressing through the various repair processes. In general, the curves are quite close.

Fig. 36 shows the state probability curves for a 20 BS network and 4 repair crews. Each curve has a different SCV value which is indicated in the legend. From Fig. 36 we can observe that as the SCV value increases, the service restoration time substantially increases.

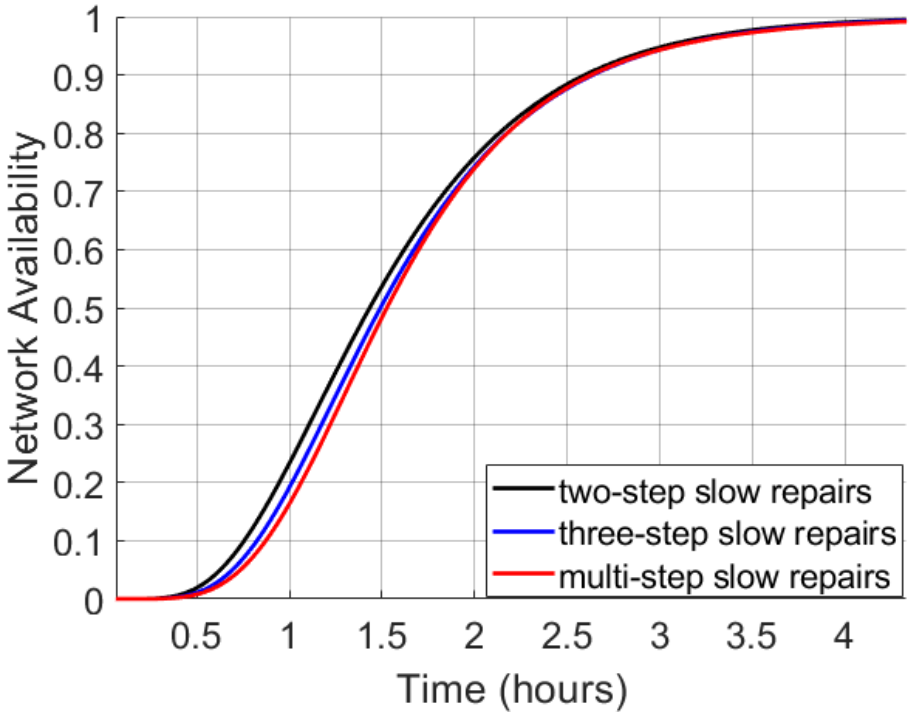

Figure 37: Transient State Probability for a BS Network of 15 BS with Three Repair Crews and Varying Non-Exponential Repair Models

3.6 Conclusion

The ME models provided here give a comprehensive understanding of the disaster recovery process with various repair models. This chapter discussed the overall formulation of the ME model with the appropriate matrix formulations, calculations of the infinitesimal rate matrix and computation of the transient probabilities including dependencies explained through the Kronecker product. The base repair model was explained in Fig. 25 and variations of this repair model were used for deriving results. The results presented the transient and steady state analysis of the ME model, provided the scalable

survivability model, discussed different aspects of survivability design, and provided the service restoration and service availability figures. The survivability design of the ME model discussed the optimal combination of the number of BSs and repair crews which would improve the restoration time and network availability time. An asymptotic approximation of higher number of repair crews has been derived. The survivability design section also talked about the balance adopted for fast and slow repairs and their effect on overall restoration time. It also included the result of changing a single parameter and determining how it effects the overall repair process. The observations made in this chapter will provide any cellular service provider with essential insight into survivable design of their network and allow them to restore the network faster while using optimal resources.

Future investigation could expand the birth-death sequential Markov chain to more elaborate models that include correlated failures of macro cell BSs, small cell BSs, and backhaul connections. Stochastic geometry could also add geographic knowledge to the model. Repairs can be prioritized to areas where little or no coverage is available when BSs fail. And in addition to the failure analysis and time dynamic understanding provided here, financial analysis could complete other survivable design considerations. It could provide cost models to optimize the combination of numbers and types of repair crews versus financial losses due to failed parts of the network.

CHAPTER 4

MULTI CONNECTIVITY BASED HANDOVER ENHANCEMENT AND ADAPTIVE FRACTIONAL PACKET DUPLICATION IN 5G CELLULAR NETWORKS

4.1 Introduction

This chapter is based on two of our papers on optimization of cellular multi-connectivity and fractional packet duplication, "Fractional packet duplication and fade duration outage probability analysis for handover enhancement in 5G cellular networks," by R. A. Paropkari, A. A. Gebremichail, and C. Beard, published in the 2019 International Conference on Computing, Networking and Communications (ICNC) and the second paper "Multi Connectivity based Adaptive Fractional Packet Duplication in Cellular Networks" by R. A. Paropkari, and C. Beard, submitted to the MDPI Signals Journal Special Issue - B5G/6G Networks: Directions and Advances. I want to thank my co-authors for their contribution towards this research and their help with writing the papers.

Handovers allow users to move around freely and connect to any nearby cell that gives them the best service, better throughput, ultra-low latency, etc. Until now, mobility was one of the prime factors to trigger a handover in any given network but several other parameters like Received Signal Strength (RSSI), Signal to Interference and Noise Ratio (SINR), Fade Duration (FD), Quality of Service (QoS), shadowing, backhaul connectivity, network congestion, reliability, etc., may also result in a handover in any given network. A handover lets a user experience good service often overcoming network la-

tency and by providing higher throughput. Also, with the need to integrate the different heterogeneous wireless networks in 5G, managing handovers will be a challenge. In the next generation of wireless networks, that includes dense cellular deployments, handovers will continue to play an important role. Any user will have to be guaranteed a continuous connection, and at the same time, too many handoffs will also have to be avoided. The 5G proposes presence of a heterogeneous environment where a user will seamlessly have to undergo a vertical handover without any service interruption times. These vertical handoffs need to be carried out seamlessly and precisely when required. Service interruption is not acceptable for certain applications, and handovers are the means to avoid loss in connectivity.

Fade Duration Outage Probability defines a time over which a communication will fail if a fade persists too long. For example, dropping successive packets can cause lost source relationships and channel coding fails after too long of sequences of errors. FDOP provides a direct relationship with quality of a connection. We show that FDOP may result in a reduced range of coverage, but also more time for the handover process. FDOP is also very helpful for multi-connectivity cases. We introduce a novel fractional packet duplication process along with FDOP to only duplicate enough packets over multiple connections to meet outage requirements.

Millimeter Wave (mmWave) frequency bands have wide available bandwidths compared to the conventional cellular frequencies. They have been of great interest and a key enabler of low latency and multi gigabit speeds for the Fifth Generation (5G) of cellular networks. The 3rd Generation Partnership Project (3GPP) introduced the New

Radio (NR) cellular standards and also included the mmWave spectrum due to the ultra-high throughput potential satisfying the enhance mobile broadband (eMBB) 5G use case requirements. The optimal use of mmWave can also help reduce the control signaling overhead and improve the overall communication latency. A lot more is possible when mmWave frequencies are used in conjunction with existing Sub-6 cellular frequencies either by means of Dual Connectivity (DC) and/or Carrier Aggregation (CA). 5G network infrastructure allows for the amalgamation of multi network convergence and due to the explosion in the number of User Equipment (UE) and Access Points (AP), carrier aggregation of of radio resources and multiple connectivity are the means to increase the coverage and capacity. We will be focused on the concepts of DC and our proposed Adaptive Fractional Packet Duplication (A-FPD) scheme throughput this paper.

The mmWave inherits several challenges of its own like the isotropic pathloss and heavy attenuation due to blockage by common materials. This makes the wireless channel extremely vulnerable to typical Non-Line of Sight (NLOS) transmission and constantly changing environmental conditions blocking the Line of Sight (LOS). In order to overcome the propagation pathloss, highly directional means of communications is implemented. Appreciating the small wavelength of the mmWave, many antennas can be packed closely together to enable Massive Multiple Input Multiple Output (mMIMO) diversity that in turn improves the link budget and range of the communication. The mMIMO is usually deployed with beam forming mechanism which enhances the directional communication ability. In order to tackle the other challenge of blockage, the Ultra Dense Networks (UDN) deployment is a method used to deploy more small cells reducing

any shadowing or no coverage zones.

4.2 Related Work

As mentioned earlier, most of the handover improvement techniques are still based upon the SNR, SINR, or the RSSI measurements. There have been some protocol enhancements also to reduce the handover signaling time. Even the use of X2 links, which connect the BSs directly, are exploited to reduce handover times and utilize backhaul resources more efficiently. The authors in [79] focus on the joint availability of power-controlled Rayleigh-fading links while using selection combining. The outage probability and the handover probability are also evaluated taking into account the effect of path loss, shadowing, Rayleigh fast fading, frequency factor reuse and conventional beamforming [80]. An outage will be caused if SINR drops below a threshold and the number of times this outage occurs decides whether a handover will be performed. In [81], a comparative study was performed on three main factors involved in handover decisions. A smart combination of RSSI, data rate, and SINR was proposed to help make a handover decision more efficient rather than using each component individually. Authors in [82] introduce a new parameter, Interference to other Interference and Noise Ratio (IINR), so that a handover was triggered only if a throughput gain existed. SINR and distance measurement parameters work together in [83] to aid handover decisions resulting in reduction of the number of handovers.

We represented handovers using matrix exponential distributions for public safety and emergency communications, which helps make handover decisions more accurate

considering all the different parameters involved in the decision process [22]. To characterize wireless channels, the authors in [84] approximate the fade duration distribution by using an exponential distribution. They use this approximation to derive the minimum duration outage of multiple selection combined links; this is a performance metric for survivability which also captures the time correlation of time-varying channels. They show that it is more efficient in terms of power to utilize multiple links in parallel rather than boosting the power of a single stand-alone link. A stochastic model of the SINR distribution captured shadow fading and more accurately characterized the SINR [85]. Instead of these, we propose the use of FDOP instead of SINR as a parameter to help make the handover decision in any cellular network. Multi-connectivity approaches with SINR evaluation were used to achieve five nines reliability [86, 87]. The authors also showed how single connectivity reliability deteriorates with mobility and proposed a multi-connectivity concept for a cloud radio access network as a solution for mobility related link failures and throughput degradation of cell-edge users. The concept relies on the fact that the transmissions from co-operating cells are coordinated for both data and control signals. The work in [88] presented a new tractable analytical framework for evaluating coverage probability in heterogeneous networks which captures the non-uniformity of these deployments.

An advanced handover scheme using the intra-frequency Dual Connectivity (DC) principle from LTE was analyzed resulting in a 0 ms interruption time and 0 call failures maintaining a reliable connection all the time [89]. FDOP and minimum duration outage metrics have been used for various applications. The authors in [90, 91] introduce min-

imum duration outages that consider durations of signal fades for channels susceptible to Rayleigh fading. Their results show that under typical Doppler frequencies, outages due to Rayleigh fading are more likely to cause frame or packet errors rather than call dropping due to the short time scales in effect. A cluster-based sleep mode activation optimization method based on FDOP was proposed in [92], along with an algorithm for hybrid femtocell networks. Multi-hop relay selection algorithms based on average fade duration (AFD) and FDOP threshold in cooperative wireless networks were proposed in [93].

Authors in [94] propose partial packet duplication to satisfy traffic reliability requirements when dual connectivity is available to provide macro diversity. Idea is to duplicate only some portion of the time as needed; this utilizes potentially much fewer resources from the secondary access point. We had introduced the concept of Fade Duration Outage Probability (FDOP) and Fractional Packet duplication in our previous work in [27] where FDOP based handover requirements were shown in contrast with the traditional SINR based handovers methods in cellular systems. Authors in [95] provide a detailed tutorial on a recently developed full-stack mmWave module integrated into the widely used open-source ns–3 simulator. The work in [96] derives new formulas for two-hop and three-hop relay paths, with 10 three hop paths given a penalty cost. Then optimization algorithms for each type of relay selection 11 method are derived, which include total path and link-by-link optimization. [97] presents an implementation for the ns-3 mmWave module of multi connectivity techniques for 3GPP New Radio at mmWave frequencies, namely Carrier Aggregation and Dual Connectivity, and discuss how they

can be integrated to increase the functionalities offered by the ns-3 mmWave module.

Transient and the steady state representations of system repair models, namely, fast and slow (i.e., crew-based) repairs for networks consisting of a multiple repair crews have been analyzed in [98] . Failures are exponentially modeled as per common practice, but ME distributions describe the more complex recovery processes. An analytical model to study the impact of handover procedures and multi-connectivity degree on the latency and reliability of blockage driven wireless networks is presented in [99]. In contrast to any traditional handover improvement scheme, authors in [100] develop a 'Deep-Mobility' model by implementing a deep learning neural network (DLNN) to manage network mobility, utilizing in-network deep learning and prediction. With highly directional beams and fast varying channels, this directional tracking may be the main bottleneck in realizing robust mmWave networks [101] and this papers deals with the mmWave space requirement for the network to track the direction of each link in addition to its power and timing. As for network intelligence, the authors in [22] represented handovers using Markov Chain Matrix Exponential (ME) distributions for public safety and emergency communications, which helps make handover decisions more accurate considering all the different parameters involved in the decision process.

The work in [102] talks about the architectural enhancements and performance analysis of Packet Duplication form URLLC in 5G. Authors in [103] do a complete in-depth survey for the Horizontal and Vertical Handovers in Heterogeneous Next Generation Wireless Networks. Authors in [104] propose an approach using New Radio Dual Connectivity (NR-DC) to maximize the throughput while ensuring ultra reliable low la-

tency communication. By combining the methods from queuing theory, stochastic geometry, as well as ray-based and system-level simulations, authors in [105] develop a novel performance evaluation methodology, considering the intricacies of mmWave radio propagation in realistic urban environments; the dynamic mmWave link blockage due to human mobility; and the multi-connectivity network behavior to preserve session continuity. In [106] an anchor-based MC mobility model has been proposed in 5G UCN environment to enhance user mobility robustness. [107] provides the first comprehensive end-to-end evaluation of handover mechanisms in mmWave cellular systems.

Multi-connectivity is explored as a solution for assuring high reliability in industrial scenarios. Several multi-connectivity techniques are compared, using real channel measurements from two factories [108]. Fog-RAN Enabled Multi-Connectivity and Multi-Cell Scheduling Framework for 5G URLLC is studied in detail in [109]. [110] describes the packet duplication functionality in 5G-NR and highlights the related technical challenges. The Survey in [111] provides an overview of different MC concepts and scheduling categories. Three main scheduling categories were identified: packet duplication, packet splitting and load balancing. A multi-connectivity concept for a cloud radio access network as a solution for mobility related link failures and throughput degradation of cell-edge users is proposed in [112]. Authors in [113] study the performance analysis of Packet Duplication for reliability enhancement of Wireless Links in 5G.

In order to provide ultra-reliable services to mobile users there is a need for network architectures that tightly and seamlessly integrate the LTE and mmWave Radio Access Technologies and two possible alternatives are presented in [114] with simulation

tools to assess and compare their performance. Authors in [115] evaluate and show the tremendous transmit power reduction of multi-connectivity over single-connectivity, by analytically deriving the corresponding SNR gain. [116] presents an analytical study of the outage probability enhancement with multi-connectivity, and analyses its cost in terms of resource usage. The performance analysis is further compared against conventional single-connectivity transmission. [117] aims to provide a broad perspective on the fundamental trade offs in URLLC, as well as the principles used in building access protocols.

4.3 Fade Duration Outage Probability Analysis for Handover Enhancement

Wireless networks are deploying a myriad of small cells for densification of the network with the aim of improving coverage and helping offload users from the macro sites. This has caused BSs to get closer to each other and users, switching connections more often between these BSs. Handover techniques have to be revised in order to avoid the ping pong effect where a user keeps switching between BSs for more than required. Many RF-related physical parameters play an important role in cellular handovers. Shadowing, multipath, channel path loss, etc. can all contribute to a handoff decision. Today, most handover decisions are based on SINR values. As the SINR drops below a certain threshold, the network initiates a handover. Especially for high speed users, the SINR changes are often quick and handover needs to happen well on time. A standard event list is defined by the 3GPP standards but the values for the threshold are normally at the discretion of the service provider and are altered as per the requirements of the user or the application.

Managing handovers in cellular systems is still one of the most important challenges we face today. Current standards and protocols make them strong but still not as efficient and reliable as required. Decisions are often designed around the SNR and/or the Reference Signal Received Power (RSRP) which have proven to be correct and have catered to the user well so far. Any enhancements or newly proposed handover techniques are also usually based on SNR/RSRP values. Handovers occur when SNR goes below a certain threshold value for more than a predetermined time period commonly known as the time-to-trigger (TTT). There often is some hysteresis (+/-) for the given threshold also to avoid the ping pong effect. Common methods of handover decisions involve having multiple sets of rules for a variety of users. For example, SNR performance is severely degraded for high speed users. So, one technique reduces the TTT and modifies the threshold values to facilitate fast handovers. Slow users would have another set of thresholds and TTT value. Unlike traditional SNR measurements, we show that handovers based on Fade Duration Outage Probability (FDOP) have a greater impact on user quality of experience.

4.3.1 Proposed System Model

In a cellular system, a user will always have one connection with the base station (BS) when it is in connected mode and when this connection changes, it is called a handover. But when the same connection with a BS changes when in idle mode, a user performs a cell selection or re-selection to get into the connected mode. Reliability of a cellular system depends on a lot of factors and the signaling and data bearer are two separate connections. This demands for multi-connectivity with different eNodeB to improve

reliability of the user connection. In our model shown above, we consider two scenarios, first with a user having a single connection with a BS, and the second has a user connected to multiple independent BSs. Each connection in the second scenario will have its own channel conditions and these connections are in parallel sending duplicate data. It is important, however, not to send fully duplicate data, since this wastes resources. We also show that this is unnecessary. We propose a fractional packet duplication strategy to just meet FDOP requirements.

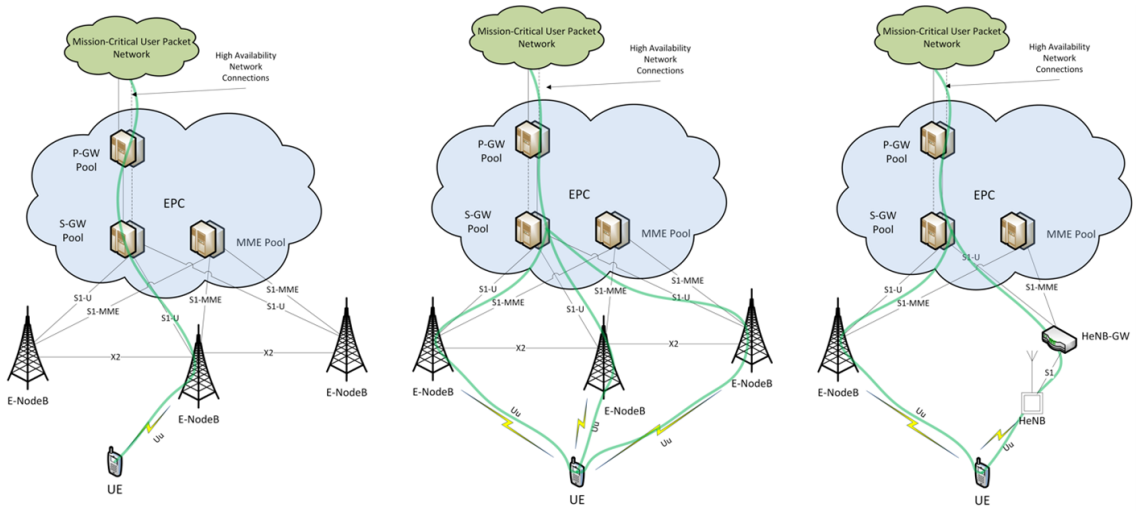

Figure 38: (a) Single Macro (b) Multiple Macro (c) Single Macro and Single Femto

Fig. 38(a) shows a user connected to a single macro BS and as it moves away from the coverage area of this BS it will be handed over to another nearby macro BS. This decision in our case will be based on the FDOP. This user is still always connected to a single BS at any given time. Fig. 38(b) shows that a user is connected to multiple macro BSs at the same time that have the lowest FDOP values. Now every independent connection is monitored and if the FDOP increases above a certain threshold value as compared to

neighboring BS, handover may be triggered, and the user will have a new connection. We ensure that the combined FDOP value is the least for all possible connections of the user at the given time instant. Say the independent $FDOP_k$ is represented as P_k where 'k' is the index of the BS. We also consider P_T as the threshold value of the FDOP for all connections in that network.

Now we have BS1, BS2 and BS3 that are connected to the user and their $FDOP_k$ are P_1, P_2 and P_3 respectively for $FDOP_{combined}$=P_1 P_2 P_3. The neighboring BS4 and BS5 have similar $FDOP_k$ based upon the user location as P_4 and P_5. When the user moves and $P_1 > P_T$ for a given time period or more, normally known as the time-to-trigger (TTT), then P_4 and P_5 will be computed and compared with the P_1 and P_T values. The combined FDOP is also compared for $FDOP_combined$=(P_1 P_2 P_3), (P_2 P_3 P_4), and (P_2 P_3 P_5). Based upon the lowest individual and combined results, the next BS is chosen which could be to choose P_4, P_5 or even stay with P_1. This helps maintain connectivity with three BS at the same time all having the lowest FDOP values and also their combined FDOP being the lowest of all the available combinations. Fig. 38(c) shows multi connectivity with a macro cell and a small cell at the same time. We represent a femto cell by HeNB or HeNodeB in our case, where H stands for heterogeneous.

4.3.2 FDOP Comparisons and Results

In this section we compare FDOP and SNR-based schemes, bit error rate (BER) and block error rate (BLER) are the two measures of error based upon the SNR values of the channel. All are compared based on outage scenarios consistent with their metrics. We show that FDOP measures require earlier handovers but with more time to complete

that handover. In the next section, through FDOP comparisons, we show the importance of multi-connected users and also introduce fractional packet duplication, that allows the best use of multiple connections. We make the common assumption for non-line-of-sight situations that the received signal is affected by Rayleigh fading. The longer it remains in a deep fade the greater the effect to the original signal. We denote the outage due to fading as FDOP, which considers the depth and duration of the fade as both significant. The FDOP between UE and BS k is the probability a signal will stay below R_0 for a duration longer than T_{thr}. In terms of f_{out},τ_{out}, FDOP can be computed as:

$$FDOP_k = f_{out}\bar{\tau}_{out} \tag{4.1}$$

where f_{out} represents the fading rate and τ_{out} is average duration of an outage. Based on Rice's work, an outage occurs when fade time τ_f exceeds the threshold value T_{thr}, the probability of which can be denoted as

$$P_{\tau_f}(\tau_f \geq T_{thr}) = \frac{2}{x} I_1\left(\frac{2}{\pi x^2}\right) \exp\left(-\frac{2}{\pi x^2}\right) \tag{4.2}$$

I_1 (x) represents a modified Bessel function order one and x =T_{thr}/AFD. The pdf of the fade duration is shown in [90] equation (18) as:

$$= \frac{1}{\bar{\tau}_f} \frac{d}{dx} c \left[\frac{2}{x} I_1\left(\frac{2}{\pi x^2}\right) \exp\left(-\frac{2}{\pi x^2}\right)\right] \tag{4.3}$$

Per [90] equation (19), the pdf of the outage duration τ_{out} (when $\tau_f \geq T_{thr}$) can be denoted:

$$f_{\tau_{out}}(\tau_{out}) = \frac{f_{\tau_f}(\tau_{out})}{P_{\tau_f}(\tau_f \geq T_{thr})} \tag{4.4}$$

From this equation and using [91] equation (5), we have simplified the fade duration outage probability $FDOP_k$ as:

$$FDOP_k = LA\left[\mathbf{e}^{\left(-\frac{2}{\pi x^2}\right)}\left(I_1\left(\frac{2}{\pi x^2}\right)I_0\left(\frac{-2}{\pi x^2}\right)\right)+1\right] \tag{4.5}$$

where

$$L = \textit{Level Crossing Rate} = \sqrt{2\pi}f_{\mathbf{m}}\boldsymbol{\rho}\mathbf{e}^{-\boldsymbol{\rho}^2} \tag{4.6}$$

$$A = AFD = \frac{\mathbf{e}^{\boldsymbol{\rho}^2}-1}{\sqrt{2\pi}f_{\mathbf{m}}\boldsymbol{\rho}} \tag{4.7}$$

$$\boldsymbol{\rho} = R_0/r_{rms} \tag{4.8}$$

and f_m=vf_c/c is the max Doppler spread for a velocity v and a carrier frequency f_c. r_{rms} is root-mean-squared of the received signal power, and R_0 is the signal amplitude threshold for a fade during which virtually all data is lost.

We consider a typical cellular network scenario with a pathloss model $|h_k|$=f_c^2 $(4\pi c)^{-2}$ d^{-n}, f_c = 1800 MHz, pathloss exponent n=3.0. We use additive white Gaussian thermal noise with the 180 kHz LTE resource block bandwidth and a minimum SNR of 13 dB. SNR is still an important parameter for packet level performance, but FDOP includes this along with fade duration and uses this to compute R_0. FDOP time threshold is used as 20 ms, which typical for VoIP packets.

Fig. 39 shows calculations of FDOP, BLER, and BER with respect to distance for our typical cellular configuration. BER is computed for un-coded 16QAM, and BLER is for the same signal using a block size of 1000 bits. For FDOP, maximum Doppler spreads of 100 Hz (16.7 m/s) and 33 Hz (5.5 m/s). Given an outage probability, the distance range that can support that performance (or better) can be seen. For outage levels up

to a little higher than 10^{-5}, FDOP has a shorter coverage range. For example, at 10^{-6}, FDOP at 33 Hz has a range of 590m, whereas BLER<10^{-6} can be satisfied up to 1075m and BER<10^{-6} up to 1260m. One could also consider a direct relationship to reliability relative to a number of "9's". A "six 9's" reliability is 99.9999%=1-10^{-6}. As a real-world example, we have collaborative research with the Federal Aviation Administration (FAA), and their National Aerospace System (NAS) engineering document has a requirement of five 9's for service availability of Safety-Critical services where loss of service increases the risk of the loss of human life [118].

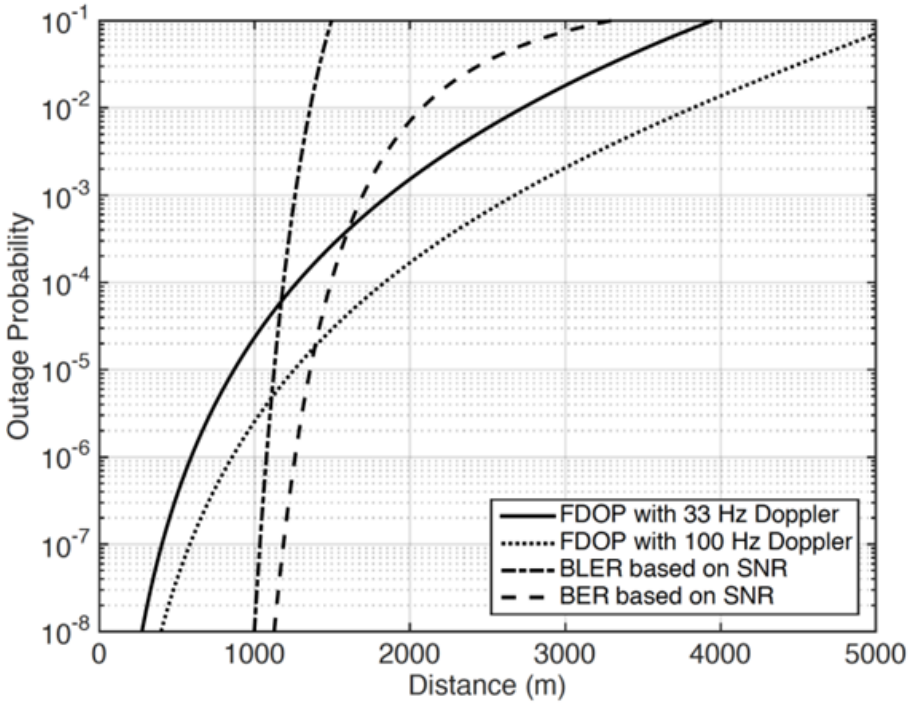

Figure 39: FDOP Outage Probability versus BLER and BER

We assume for the next figures that cell coverage area was set at a radius of 1075 m where BLER= 10^{-6}. Fig. 40(a) shows the coverage areas for 33 Hz FDOP relative to

that cell. Rings represent outage probabilities 10^{-6},10^{-5},10^{-4} and 10^{-3} outward from the center. Fig. 40(b) shows coverage for BLER with rings at those same thresholds. While FDOP cannot cover 10^{-6} and 10^{-5} outage levels as far, the coverage areas for the other outage levels extend much farther than for BLER.

There are several implications from these plots for handover processes.

- For stringent outage probabilities, such as 10^{-6} and 10^{-5} in our examples. FDOP handovers will be needed earlier.

- FDOP coverage is actually better for higher velocity UEs. This is somewhat counterintuitive, but the stems from the fact that fades are shorter duration.

- Handover processes based on SNR may be missing the impact of BLER. Once the BLER reaches the threshold, the performance gets worse very quickly.

- FDOP requires measurement of Doppler spread in addition to SNR. This will be discussed in a sequel, along with a full investigation of T_{thr} and R_0 for different applications.

- Since FDOP has more implications for performance of applications and for user expectations, we can take advantage of the fact that FDOP degradation occurs more slowly. This actually provides more time for handovers than may have been assumed previously. Some hysteresis (+/-) here for the given threshold also avoids the ping pong effect.

- Since FDOP decays slowly, it also provides more opportunities for dual connectiv-

ity, discussed in the next section.

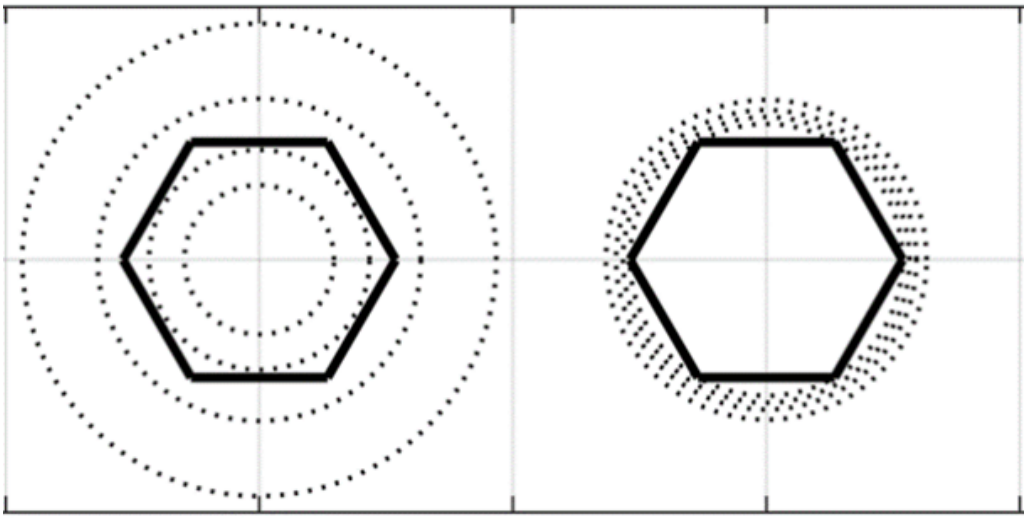

Figure 40: Coverage Range based on (a) FDOP and (b) BLER

4.3.3 Multi-Connectivity and Fractional Packet Duplication

In Fig. 41 we also show the implications of diversity using more than one parallel connection between a user and BSs. We have plots for a single receiving connection (1r), two connections (2r) and three parallel connections (3r) for 33 Hz. Diversity branches here have the same average received power. The figure also illustrates a single and two connection diversity for 100 Hz Doppler shift. It can be very clearly seen from the figure that the reliability is much higher for multiple connections and the FDOP reliabilities stay low even when the distance from the BS increases. 3GPP Release 12 introduced into LTE.

An important issue, is how would dual connectivity be implemented? Would all traffic be fully duplicated over the multiple links? In 3GPP Release 14, signaling protocols in support of packet duplication control were standardized [102]. Little research has been

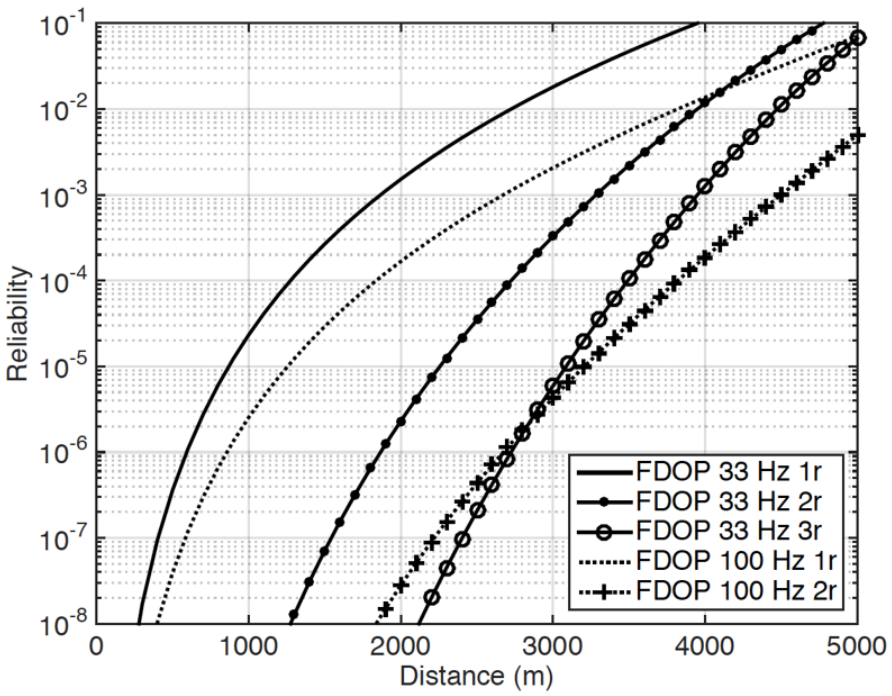

Figure 41: FDOP with 1, 2 and 3 Parallel Connections

conducted on use of this feature, so we provide insight based on FDOP. Fig. 42 shows a scenario with two base stations at distances d=0 and 2000 m. A mobile is moving from BS_1 to BS_2 with a 10^{-6} FDOP requirement. The "Only BS_1" curve shows how the FDOP gets worse away from BS_1 and "Only BS_2" shows how it improves while approaching BS_2. After 590 m, it is necessary to connect to both BS's to have an FDOP less than 10^{-6}. For example, at 750 m, only BS_1 would give $FDOP_1$ = 4.2 × 10^{-6} and only BS_2 with $FDOP_2$ = 8.7 × 10^{-5}. If the 2nd link duplicates all packets, the combined FDOP would be $FDOP_1$ × $FDOP_2$ = 3.6 × 10^{-10}. This is much below the requirement, and full duplication is wasteful of network resources. Instead, we can show that if only 75.9% of the packets are duplicated on the weaker link (what we call fractional

packet duplication), FDOP = 1×10^{-6}. The solid line shows the result of fractional packet duplication; the FDOP requirement is met at all distances. Fig. 43 shows the proportion of packets duplicated at each distance. Fractional PD is used between 590 m and 1410 m. Resource usage is significantly reduced, especially when only a few packets are duplicated. Only from 870 to 1130 m are > 90% duplication rates required.

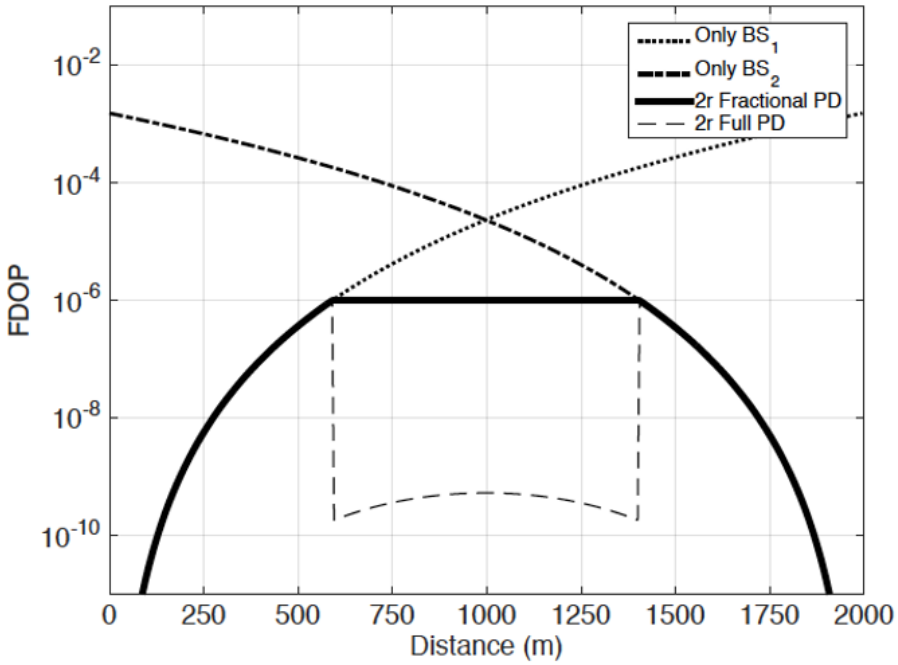

Figure 42: Reliability for Two Base Stations with and without Fractional Packet Duplication

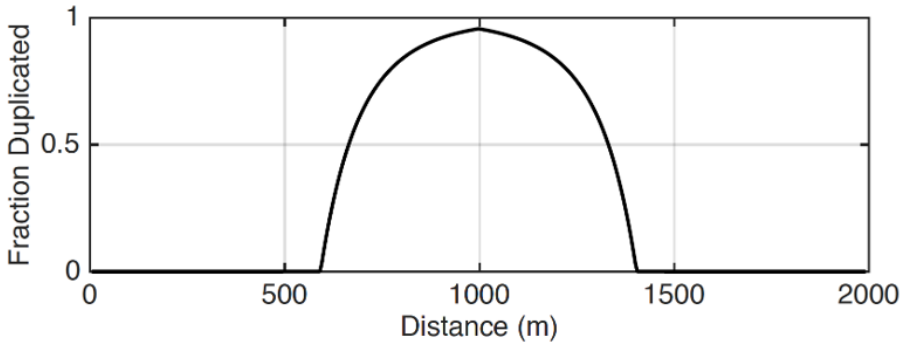

Figure 43: Fractional Packet Duplication Proportions

4.4 Multi Connectivity and Adaptive Fractional Packet Duplication in 5G

Any end User Equipment (UE) in the form of a cell phone, mobile tablet, laptop computer, mobile hotspot, wireless sensor, etc. is considered to have Multi Connectivity (MC) when connected to more than one Base Station (BS) simultaneously. Most often, this is termed as a Dual-Connectivity (DC) with just two main connections to two BS at a time. These two independent RF connections could be between the BS of the same technology or between two different technology BS like the LTE, 5G NR, UMTS, WiFi, etc. in conjunction with the Multi-Radio Access Technology (MRAT) standards. MC is also identified an a critical URLLC enabler as it adopts Spatial diversity where more than one connection serves the UE from geographically split locations. In addition, Time and Frequency diversity are also integral to MC. Time diversity could be achieved by the means of re-transmissions and error correction methods adhering to the delivery within the expected time interval. The frequencies of multiple signals are separated by coherence bandwidth if on the same band or multiple frequencies are combined for the same transmission. This can also be achieved by the means of carrier aggregation where data is

separated over different fading channels. MC has also contributed for the transition from existing infrastructure to the new 5G networks.

4.4.1 Carrier Aggregation

Introduced in the Advanced Long Term Evolution (LTE-A) standard, Carrier Aggregation (CA) is a means to combine together two or more carrier components (CC) to increase the transmission bandwidth capacity. This increment in capacity can be achieved on the DL and the UL, usually the former being higher than the latter. The concept of carrier aggregation was first introduced in Rel 10 of the 3GPP where a maximum of 5 carrier components was allowed on the downlink channel. Since then, this concept has evolved and allows multiple CA capabilities across different technologies. CA can be inter-band, meaning CC aggregation between different frequency bands, or can be intra-band, which means CC aggregation within the same frequency band. The intra-band is further categorized either as a contiguous or a non-contiguous CA which is explained below. The CA is technically a MAC-layer split and is implemented at the physical layer.

- Inter-Band CA: As shown in Fig. 3(a), CCs from different frequency bands are combined together.
- Intra-Band Contiguous CA: As shown in Fig. 3(b), CCs from the same frequency band, which are adjacent to each other, are combined together.
- Intra-Band Non-Contiguous CA: As shown in Fig. 3(c), CCs from the same frequency band which are non-adjacent or fairly spaced apart in frequency domain, are combined together.

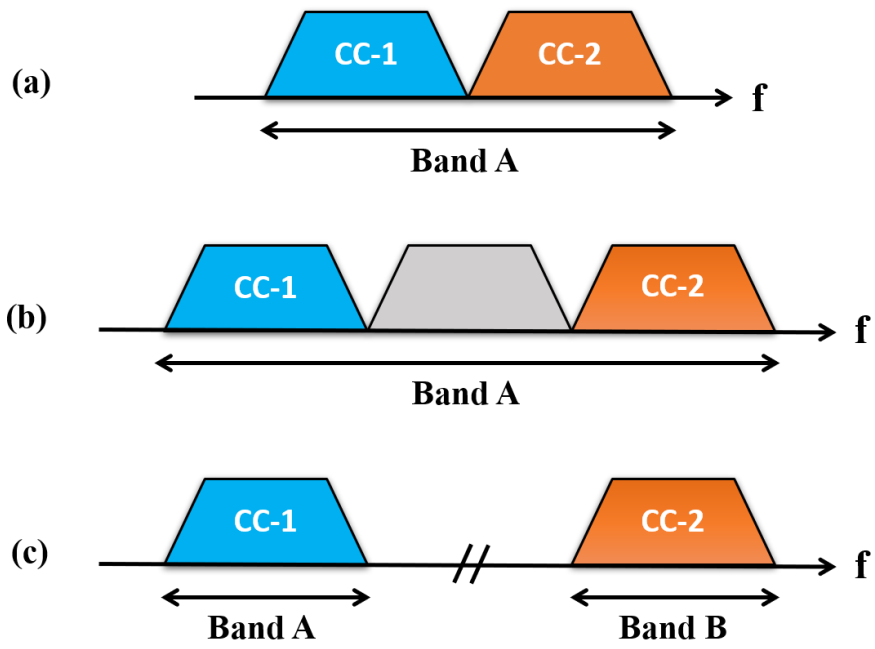

Figure 44: (a) Intra-Band Contiguous (b) Intra-Band Non-Contiguous (c) Inter-Band CA

4.4.2 Dual Connectivity

Dual Connectivity (DC) introduced in the Rel 12 of 3GPP specifications allows a UE to be simultaneously connected to two different BS operating on different frequencies. CA usually uses the radio resources of the same BS and same technology but is always limited by the scarcity of bandwidth availability. DC on the other hand allows the mobile operators to use the abundant bandwidth resources from different BS to improve the overall user experience. DC is responsible to increase the user throughput, improve mobility robustness and also improve resiliency with added diversity. DC is the prime factor resulting in speedy deployment of the 5G wireless networks worldwide. DC fuels the air interface design improvements and helps satisfy the stringent latency and reliabil-

ity requirements of the new 5G use cases. The faster emergence of the 5G is due to the most accepted approach of deployment Option 3 as defined by the 3GPP standard specification. This is further categorized as option 3a and 3x based on the user plane diversity from the UE to the core network.

The CA feature is also deployed in addition with the DC concept, including those in the Multiple Radio Access Technology (MRAT) environments. For examples, CA of one LTE CC and one NR CC leads to E-UTRAN New Radio Dual Connectivity (ENDC) at higher layers and an ultimate CA at the physical layer. DC can also be purely LTE base or NR based, like the NRDC solution involving one Sub-6 NR gNB and one mmWave 5G gNB with a PDCP split. The Figure shows Single-RAT and MRAT deployments with CA feature enablement. With DC, a UE is simultaneously connected to two different base stations: a master eNB (MeNB) and a secondary eNB (SeNB). The MeNB and the SeNB operate on different carrier frequencies. The groups of serving cells associated with the MeNB and the SeNB are referred to as the master cell group (MCG) and secondary cell group (SCG), respectively. DC is only applicable to UEs in RRC connected mode.

4.4.3 Packet Duplication

Packet Duplication is implemented at the Packet Data Convergence Protocol (PDCP) layer and can be done for both the control and the data plane. With a UE having a DC, the source node is responsible to duplicate packets and send over the two independent networks. These are then combined at the receiver with duplicate ones discarded. Packet duplication in DC can be implemented with minimal impact via the split bearer architecture, which can also be seen in Fig.45. The PD is similar to the split bearer operation with

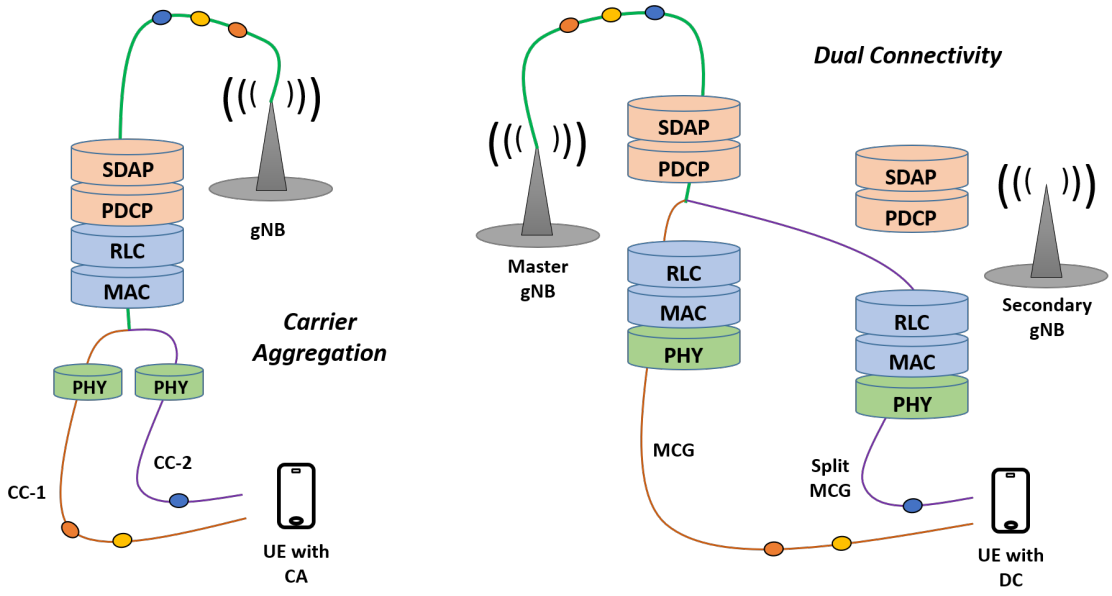

Figure 45: Carrier Aggregation and Dual Connectivity

the same PDCP Packet Data Unit (PDU) being sent over the two separate Radio Link Control (RLC) / Medium Access Control (MAC) entities or the two nodes.

The PD operation is configured by the Radio Resource Configuration (RRC) layer and usually done at the radio bearer level. When duplication is configured for a radio bearer by RRC signaling, an additional RLC entity and an additional logical channel are added to the radio bearer to handle the duplicated PDCP PDUs. In the case of DC, the two legs belong to different cell groups, MCG and SCG. Packet duplication may not always be beneficial during a bearer's lifetime. So, it should basically depend on channel conditions and the state of the radio bearer. Also, this is a very wasteful operation for the radio resource. Hence, dynamic control of packet duplication is desired; packet duplication must be dynamically activated or deactivated. The dynamic activation/deactivation of packet duplication operation avoids unnecessary waste of air interface resources.

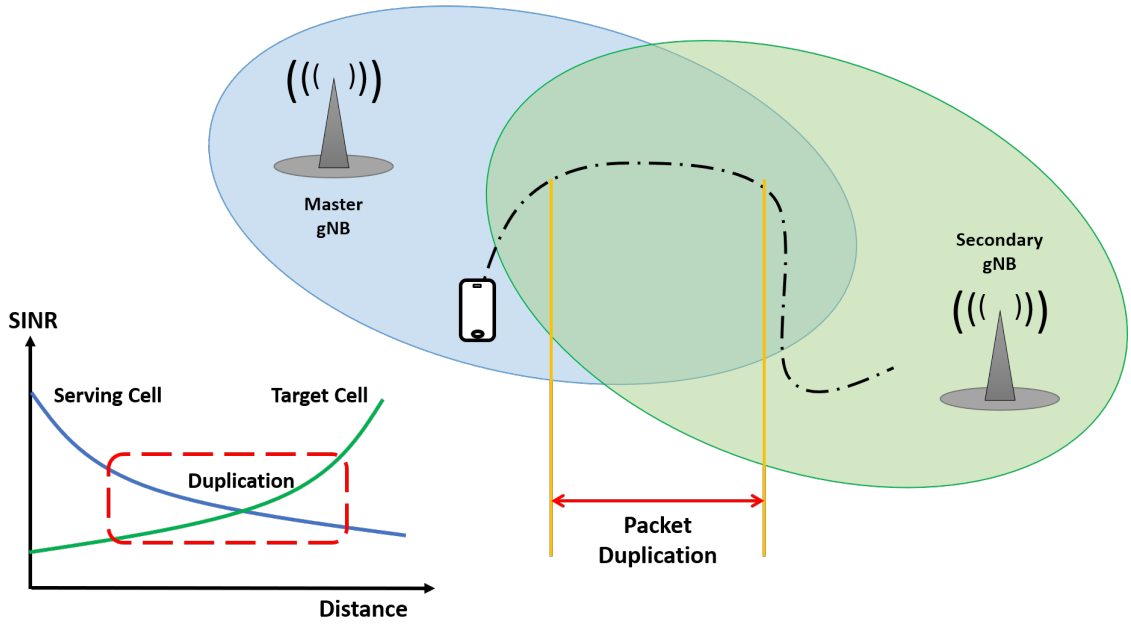

Figure 46: Adaptive Packet Duplication

Unlike DC, in CA user data is split in multiple carriers at the MAC layer. Similar to the DC case, packet duplication is configured by the RRC layer. When duplication is configured for a radio bearer by RRC, an additional RLC entity and an additional logical channel are added to the original RLC entity and the logical channel pertaining to a radio bearer to handle the duplicated PDCP PDUs. However, there is a single MAC entity as opposed to two separate MAC entities in the case of DC. It has been agreed upon in 3GPP RAN2 that PDCP duplication on the same carrier is not supported. Therefore, unlike the DC case, the mapping of the original and duplicate logical channels to different carriers also needs to be configured by the RRC layer. 3GPP RAN2 has agreed that packet duplication in CA is not supported if it is already configured in DC.

It is noteworthy that the PDCP layer in LTE already supports duplicate detection

functionality based on the sequence number. Therefore, if the transmitter sends duplicate PDCP PDUs (via different legs), only the earlier received PDCP PDU can be processed at the receiver. The PDCP PDU arriving later is simply discarded without requiring any changes in the specification. Hence, packet duplication can also be extended to the LTE-NR DC scenario.

4.4.4 Advantages of Multi-Connectivity

In this section, we briefly discuss the advantages of using the MC/DC.

4.4.4.1 Enhanced Throughput

The UE will be receiving communication over two independent RF links and this can be fully utilized to sum up the data on both links to obtain higher throughput. In Ideal conditions, this will be the total theoretical value addition of the two independent throughput, however, the channel and subsequent RF conditions always have a negative impact. A challenge is often related with the delay difference between both RF paths or the out-of-order arrival of packets at the destination which can affect the performance of upper layers, indeed reducing the throughput.

4.4.4.2 Improved Reliability

Wireless medium is often termed to be a lossy medium and re-transmissions usually make up for the reliability of wireless communications. This in fact is time consuming and utilizes the rare radio resources which not only affects the latency requirements but also negatively impact the data transmission on the radio links. Using MC, the re-transmissions can be reduced as packets can be sent over two channels simultaneously

meeting the low latency requirements. Spatial diversity also adds up to the reliability by reducing packet loss and error correction.

4.4.4.3 Robust Mobility

With MC (or DC), UE is connected to both the BS at the same time. This allows for a simultaneous control and/or user plan connectivity over two independent radio channels. DC can help reduce the interruption times during the handovers along with the amount of control signaling required. The control signalling is either already established on the secondary BS or can be moved along easily since UE has UL and DL with the primary BS. MC can help offload the overhead signalling from the core network to the Radio Access Network (RAN) due to the existing secondary node connection.

4.4.4.4 Deployment Savings

With the advancement of wireless communications and the increasing number of devices requiring extremely reliable and high bandwidth connections, service providers are always trying to improve their network's coverage and capacity. This includes deploying more BS and utilizing resources from different technologies. The Operational Expense (OPEX) is very high and a means to help transition to 5G networks is by implementing Dual Connectivity. This allows for existing 4G/LTE BS to work in conjunction with newer 5G-NR BS to provide better user experience. Also, replacing the existing infrastructure takes many years and MC allows for the progressive conversion to newer technology without service interruption.

4.4.5 Limitations of Multi-Connectivity

In this section, we describe the challenges encountered in the MC operation.

4.4.5.1 Delay and Packet Reordering

Since UE is connected to two different RATs, the Radio Resource Management (RRM) procedures can be different and radio link conditions can also add up to the transmission delay. The packets might very well arrive out-of-order at the UE. A proper packet reordering mechanism is needed to solve this problem and avoid excessive buffering which leads to degraded services for time sensitive applications.

4.4.5.2 Cross Layer Design

This is critical in MC as this can cause to a complete failure of achieving the primary goals. Proper information sharing is required to achieve efficient usage of network resources and blend in flexibility. Protocol layers are different with different technologies and have unique abilities and functionalities. All network resources have to optimally utilized and designing cross layer is a challenge given the multiple factors affecting the transmission over wireless channels.

4.4.5.3 Management of Multi-Connectivity

Networks are becoming intelligent with the evolution of Software Defined Networks (SDN) and (Network Function Virtualization) NFV but their adaption to existing cellular networks will take time. Currently, almost all of the network operators make this decision manually based on their network Key Performance Indicators (KPI). However,

the environmental conditions change and to incorporate these manually into network's decision making is almost impossible. Incorrect decisions on when to activate MC and when to use SC can degrade user experience. Such decisions can be improved by using the reinforcement learning and data analytic techniques applied to the network KPI.

4.4.6 ns-3 and mmWave Module

The Network Simulator 3 (ns-3) is an open source platform enabling the simulations of multiple different protocols for cross-layer design and analysis. Based on the already established LTE LENA platform, ns-3 has come up with a new mmWave module that is highly modular and flexible which help researchers design and validate their work. This is a full stack implementation with multiple examples and a wide variety of test configurations, all designed using C++ [95]. We make use of the Dual Connectivity (DC) functionality on the mmWave module. We utilize the MATLAB tool to further simulate packet traces received on the Downlink (DL) and the Uplink (UL) to reduce the computational overhead on ns-3. Our MATLAB code is used to precisely determine the amount of packet duplication required to maintain a certain Quality of Service (QoS) given the application. We optimally can turn ON and OFF the packet duplication in the environment based upon our scenarios described in further sections.

Fig. 47 gives a high level representation of our simulation layout. We have a User Equipment (UE) which is dual stack capable, meaning that it supports LTE as well as 5G mmWave, and is moving from point A to point B. We make use of Dual Connectivity using two Base Stations BS-1 and BS-2. There is a building between the UE and the two bases stations. At point A, UE will have some SINR received from both BS but it is

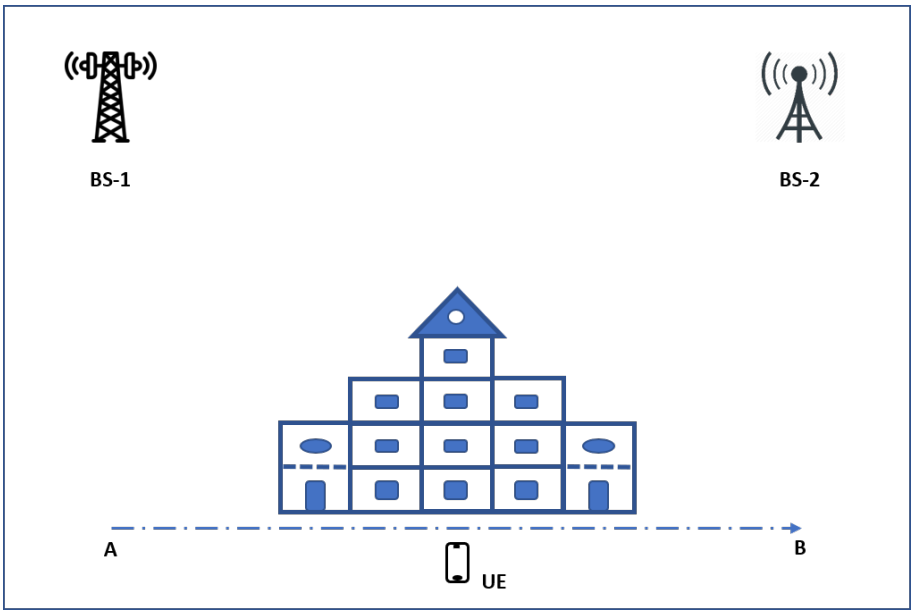

Figure 47: ns-3 Simulation Setup with One UE, Two Base Stations and One Building

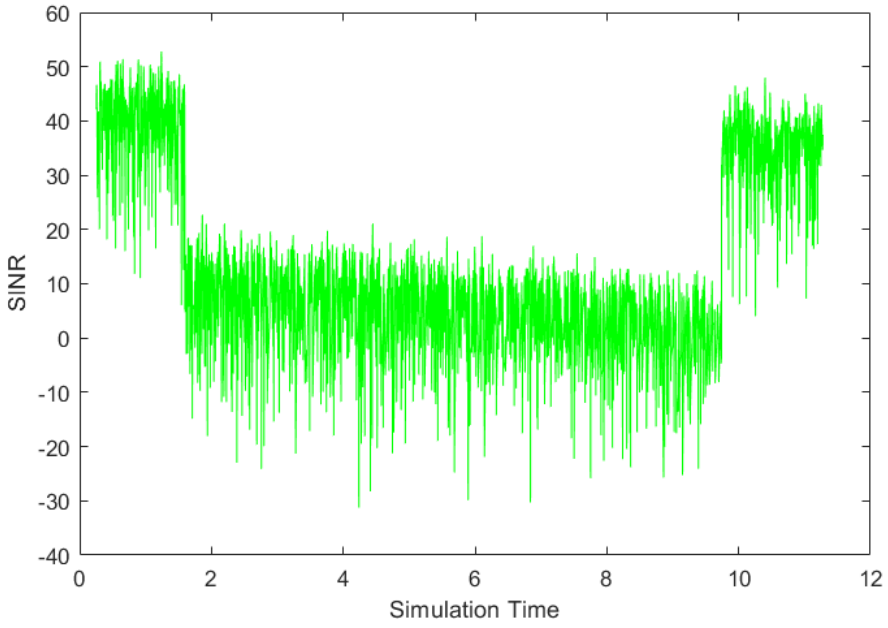

Figure 48: Instantaneous SINR of Base Station 1

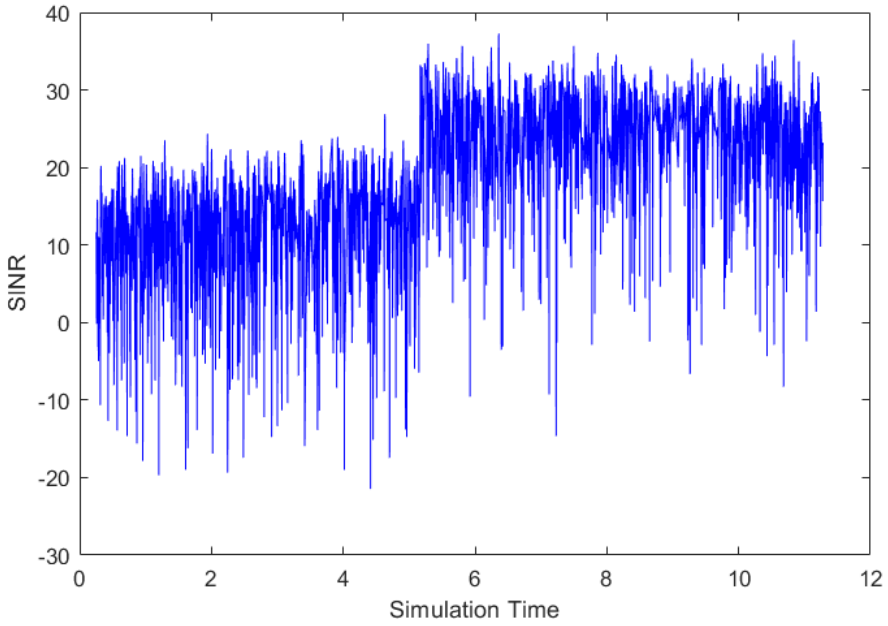

Figure 49: Instantaneous SINR of Base Station 2

closer to and has a line-of-sight (LOS) with BS-1, so the SINR from BS-1 is stronger. As it moves, this SINR is reduced when UE is behind the building, and this is where is gets closer to BS-2 and now both the BS SINR are moderately lower. This region behind the building is where our UE has a non-line-of-sight (NLOS) with both the BS-1 and BS-2. Finally, as the UE crosses over the building and has a LOS with BS-2, the BS-2 SINR gets better. This is also when a LOS is established again with BS-1 improving the SINR. Our UE is always connected to the two BS and this means that the user plane connectivity is always enabled on both the RF links of the two BS. MC (or DC in our case) represents that the UE in connected mode is configured to use the available radio resources of both the BS. So, in the case of a radio link degradation on one of the BS, the other radio link can be used for the data transmission. This helps with the Radio Link Failures (RLF) and

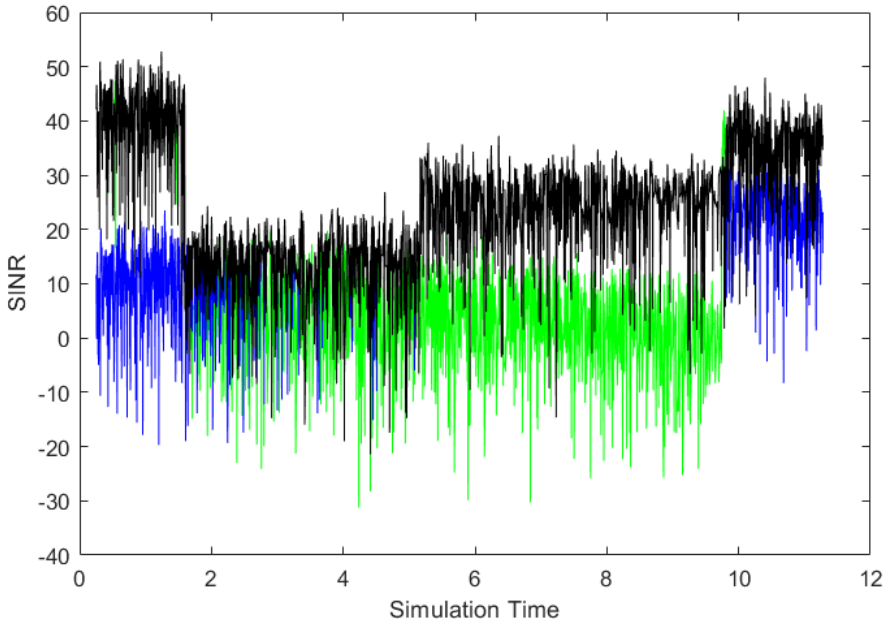

Figure 50: Instantaneous and the Best SINR of the Two Base Stations

service disruptions to be significantly reduced.

Our UE is continuously measuring and reporting the SINR of both the Base Stations. Fig. 48 shows the SINR of BS-1 received on the UE and Fig. 49 shows SINR of BS-2. We compare and select the better of the two signals at every instant and we represent that as the Best SINR. Fig. 50 has the Best SINR and the instantaneous SINR from the two base stations. The SINR for BS-1 varies from 50dB to -30dB whereas the SINR for BS-2 varies between 35dB to -20dB and the Best SINR will pick the better signal of the two. The UE sends and receives data from both the base stations on the UL and DL. Not all data received on the UE is usable as some of the packets could be corrupted and some could be completely lost due to a deep fades in the signal. So, we make sure that even if the SINR is acceptable, the data received is not corrupted. We do this by dupli-

cating the data packets for that particular corrupted packets. Either of the two BS will be used to duplicate packets based on its user connectivity with the UE.

We have discussed some of the Pros and Cons of Packet Duplication (PD) in earlier sections. Always ON PD is a wasteful utilization of the available resource and so we proposed Adaptive Fractional Packet Duplication (A-FPD) which will adapt to the channel conditions and duplicate packets only when necessary on the secondary RF link. We proposed multiple schemes to turn ON and OFF the Packet Duplication (PD). The first scheme proposed for PD is when the SINR difference between the two base stations is under a certain predefined threshold value, called Delta. A smaller SINR difference, or Delta, means that the channel conditions for the two base stations are similar and a higher difference means that the RF channel conditions are very different for the two base stations. Our goal is utilizing both the RF links when SINR received from both base stations is similar, or the delta is small. So, we will turn ON PD for a smaller delta threshold value and turn it OFF when delta is off the threshold limit. Duplicating packets with a higher delta will not benefit as much since one of the SINR value will be worse than the other and anything received on this worse link will be corrupt and UE will always chose the packets received on the better SINR link.

4.4.7 Simulation Results

As mentioned earlier, we make use of the ns-3 network simulator and especially the mmWave module of the simulator that is being developed by NYU Wireless and the University of Padova [95]. This module is specifically designed for the simulation of 5G cellular networks operating at mmWaves. It has custom PHY and MAC classes to sup-

port the 5G NR frame structure and the numerologies. It supports Carrier Aggregation (CA) at the MAC layer and also supports Dual Connectivity (DC) with LTE BS. We have approached our Adaptive-Fractional Packet Duplication schemes in 3 different ways as mentioned below in detail. SINR Threshold or Delta SINR is using two or more RF channels to duplicate packets when their RF characteristics are not very different from each other. The second method uses the Fade threshold where if a signal drops below a certain value, packets will be duplicated on two or more RF links. The third method is Distribution based where our rate of packet duplication depends on the random exponential variable.

4.4.7.1 SINR Based Packet Duplication

The activation and deactivation of packet duplication requires control signalling and if this is done many times, a lot of radio resources will be used for the control signalling which will be against our goal of efficient utilization of RF resources. If the instantaneous SINR is to be taken into account to make decisions on PD, we observed that the activation-deactivation operation happens multiple times over a single data communication session. So, we average out the SINR over a certain sample size, and then use the Average SINR value for PD. This helps reduce the number of switches and hence the signaling overhead. We have shown two sample size, 500 and 50, to average out the instantaneous SINR from both the BS. PD activation and deactivation for multiple delta threshold values for a average SINR sample size of 500 is shown in Figs. 51 - 52. Similarly, the same is shown in Figs. 53 - 54 for average SINR sample size of 50, showing many more on-off PD transitions.

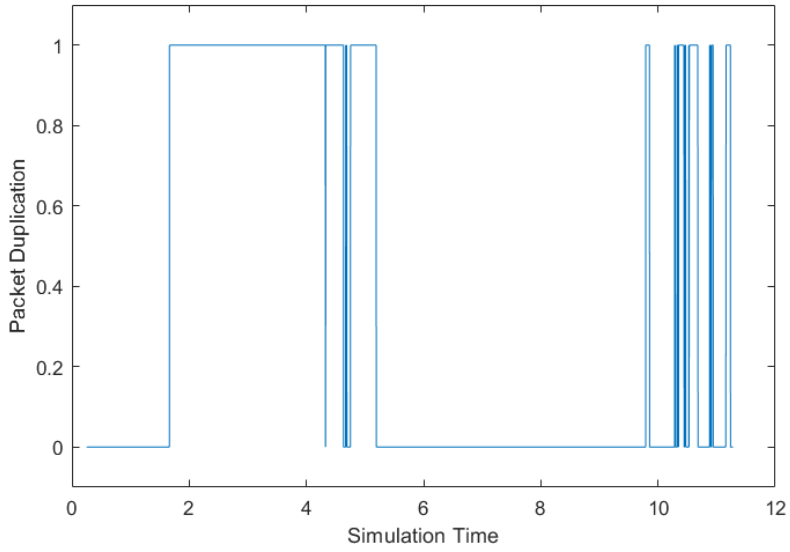

Figure 51: PD with Average (500 Sample Size) SINR difference of <=10dB

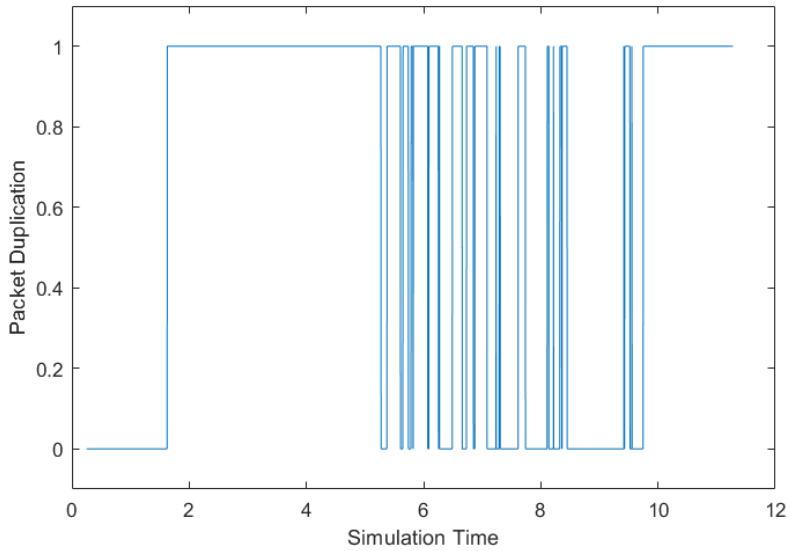

Figure 52: PD with Average (500 Sample Size) SINR difference of <=20dB

Regarding the difference in SINR values, the smaller the delta threshold, less likely is the PD as RF channel conditions for the two base stations are different. And

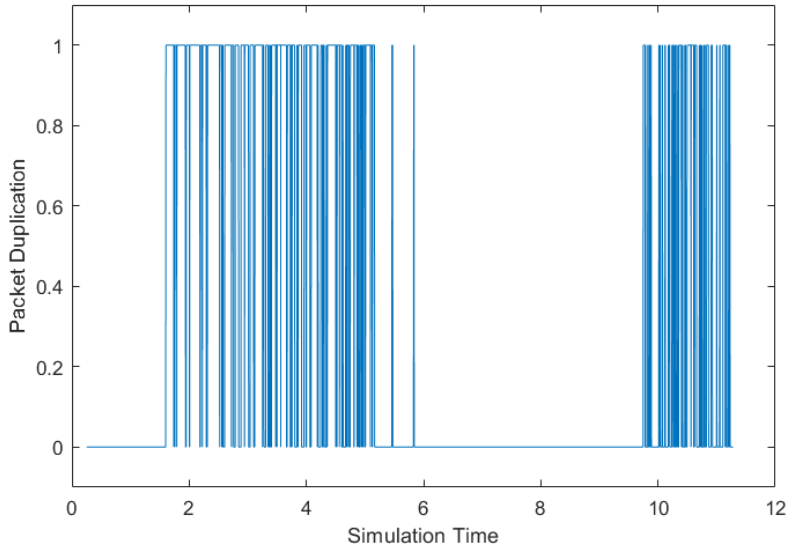

Figure 53: PD with Average (50 Sample Size) SINR difference of <=10dB

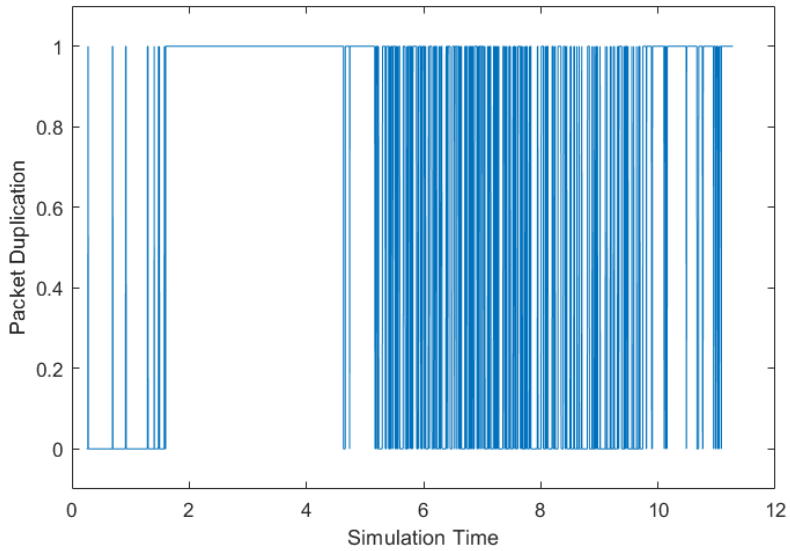

Figure 54: PD with Average (50 Sample Size) SINR difference of <=20dB

a higher delta threshold means more PD. You will notice that the PD will not toggle more often with higher delta but will have more switching for a smaller delta threshold, clearly

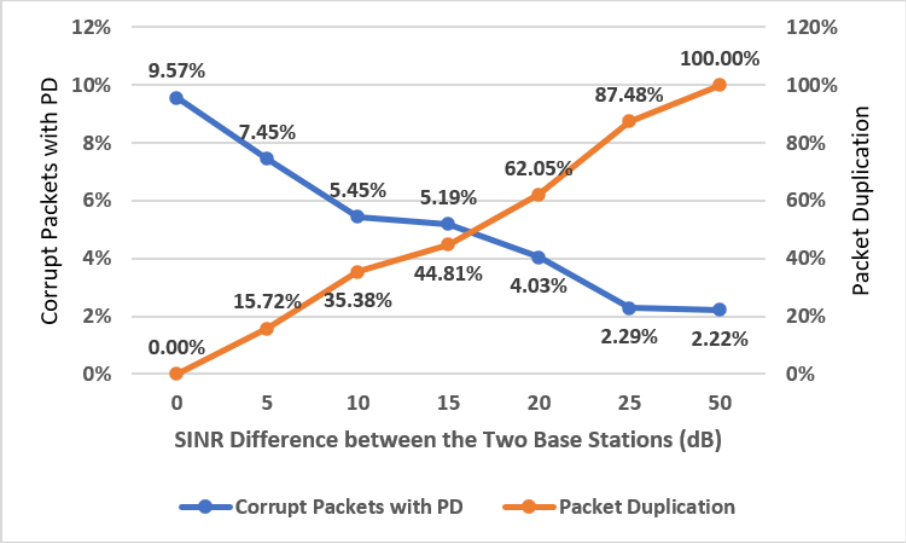

Figure 55: Difference in Average SINR (500 Sample Size) Based PD

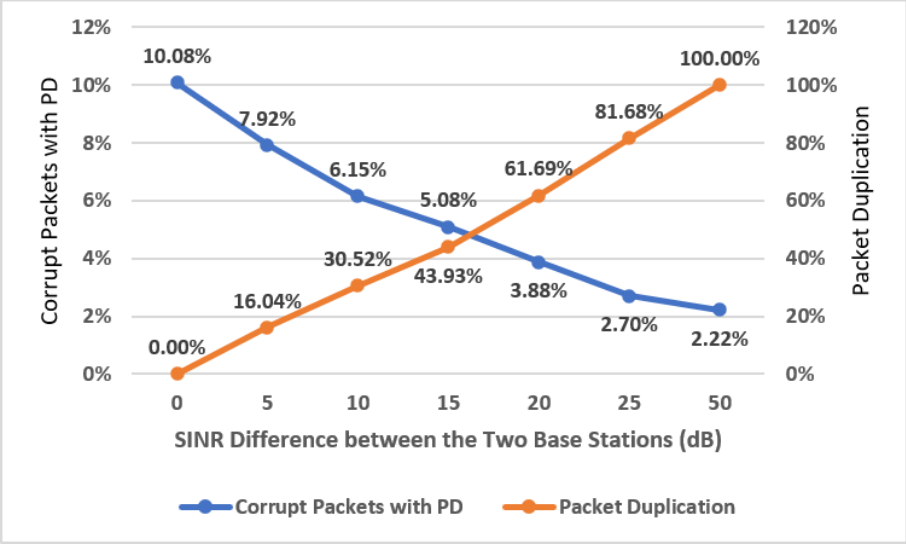

Figure 56: Difference in Average SINR (50 Sample Size) Based PD

showing that even average SINR has many fluctuations over time given the unpredictable RF conditions. Fig. 55 shows for the delta SINR on the X-axis, how much reduction in corrupt packets can be achieved with PD. We also show how much of the actual packet

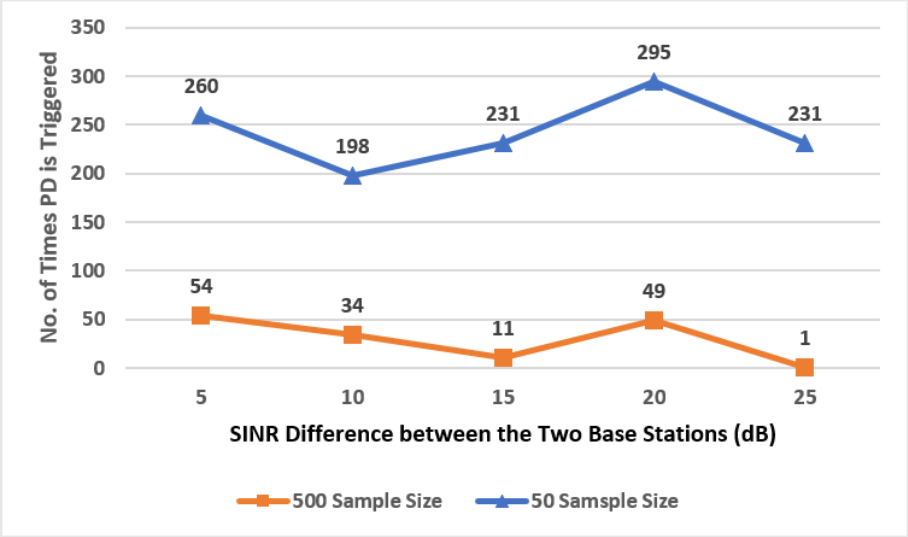

Figure 57: Actual Number of Times PD is Triggered

duplication will be required to reach this number. For example, 35.38% packet duplication will be required to have 5.45% of corrupt packets in the overall communication. This amount of PD will increase in order to achieve minimum corrupt packets. In contrast to the average SINR with sample size of 500, we simulate the environment using average SINR with sample size of 50 and the chart is shown in Fig. 56. Average SINR with 50 sample size has a lot more fluctuations than the average SINR with 500 sample size, the PD switching happens a lot more times. Fig. 57 will show you the number of times packets duplication was triggered for average SINR with 500 and 50 sample size for the two base stations.

4.4.7.2 FDOP Based Packet Duplication

As discussed in the earlier portion of this chapter, Fade Duration Outage Probability (FDOP) defines a time over which a communication will fail if a fade persists too

long. As per Fig. 48 and Fig. 49, the average SINR with a smaller sample size of 50 shows very drastic changes over a very small interval of time. If the SINR falls below a certain minimum acceptable value, any packets transmitted over that time interval could be either corrupt or completely lost. This fading of the signal below a certain threshold value is used to decide whether or not the packets will be duplicated. Figs. 58 - 59 show the packet duplication operation for the different Fade Threshold values used in our simulation for average SINR sample size of 500. Figs. 60 - 61 represent the very same information for average SINR sample size of 50.

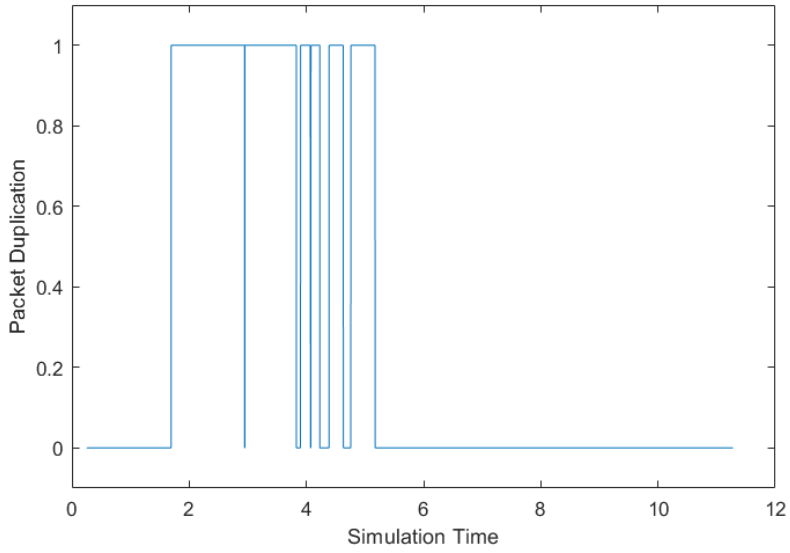

Figure 58: PD with Average (500 Sample Size) Fade Threshold of 15dB

Fig. 62 shows for the Fade threshold on the X-axis, how much reduction in corrupt packets can be achieved with PD. We also show how much of the actual packet duplication will be required to reach this number. For example, 52.52% packet duplication will be required to have 4.05% of corrupt packets in the overall communication. This amount

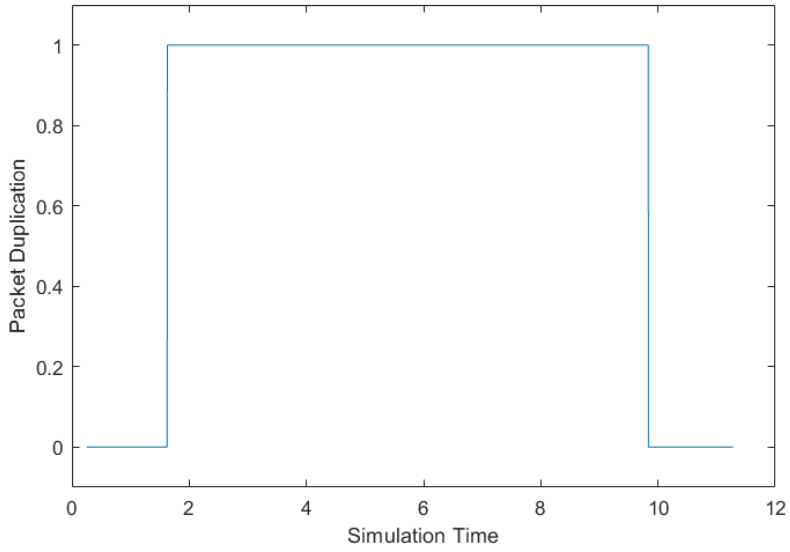

Figure 59: PPD with Average (500 Sample Size) Fade Threshold of 30dB

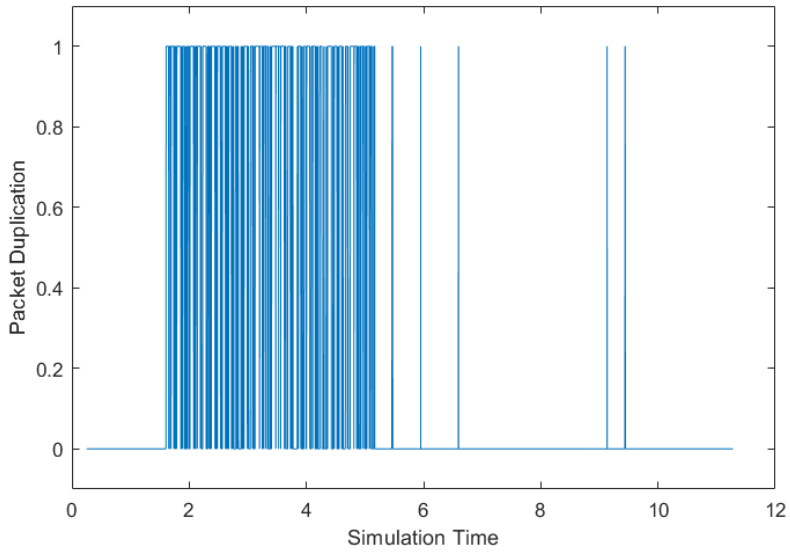

Figure 60: PPD with Average (50 Sample Size) Fade Threshold of 15dB

of PD will increase in order to achieve minimum corrupt packets. This also means more radio resource utilization. In contrast to the average SINR with sample size of 500, we

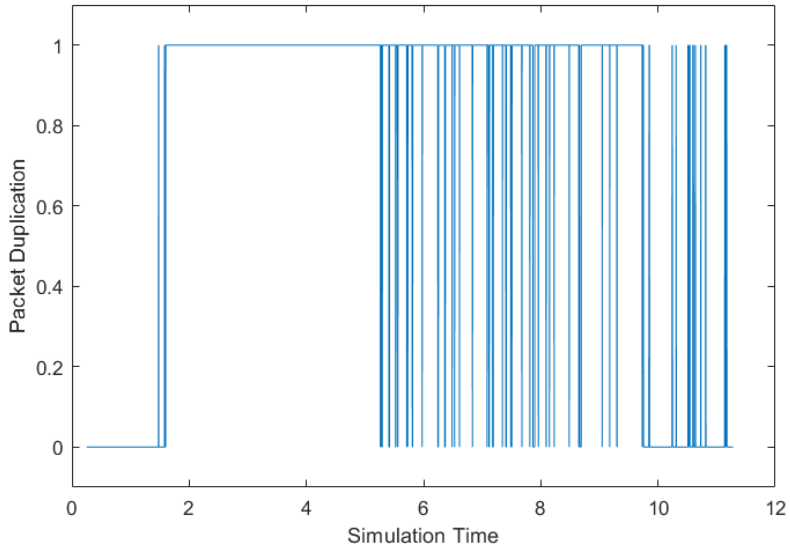

Figure 61: PD with Average (50 Sample Size) Fade Threshold of 30dB

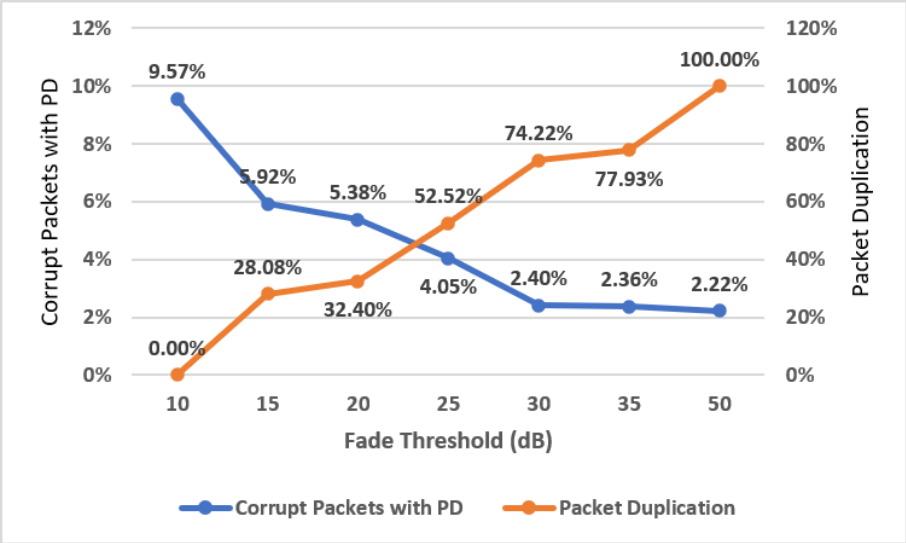

Figure 62: Fade Threshold (500 Sample Size) Based PD

simulate the environment using average SINR with sample size of 50 and the chart is shown in Fig. 63. Average SINR with 50 sample size has a lot more fluctuations than the

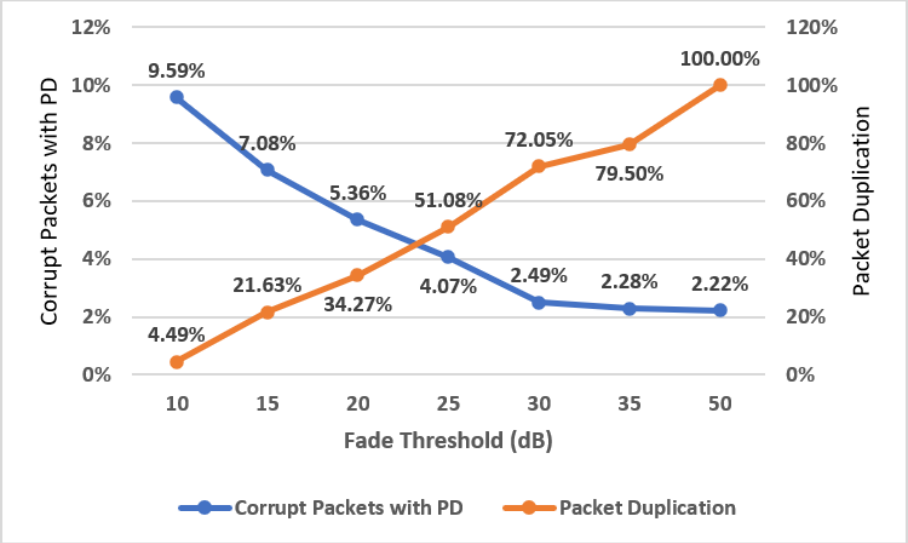

Figure 63: Fade Threshold (50 Sample Size) Based PD

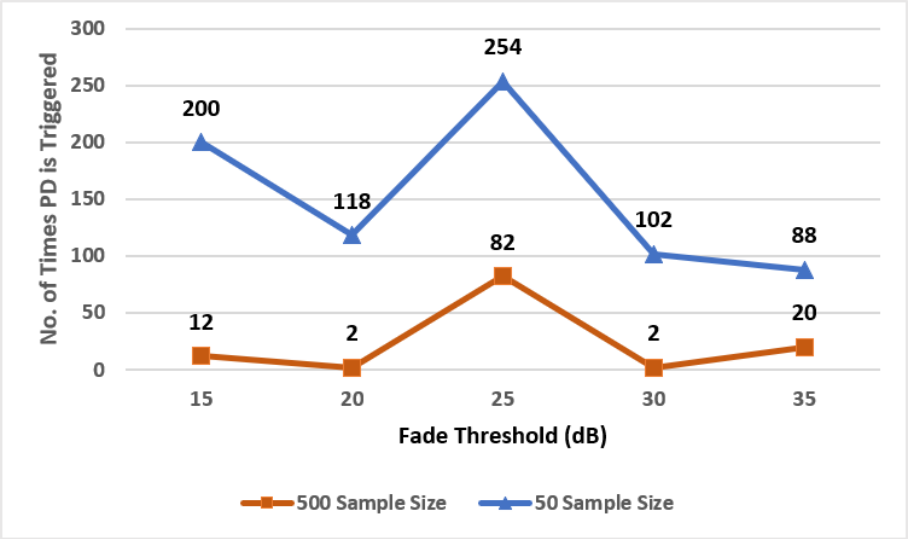

Figure 64: Actual Number of Times PD is Triggered

average SINR with 500 sample size, the PD switching happens a lot more times. Fig. 64 will show you the number of times packets duplication was triggered for average SINR with 500 and 50 sample size for the two base stations.

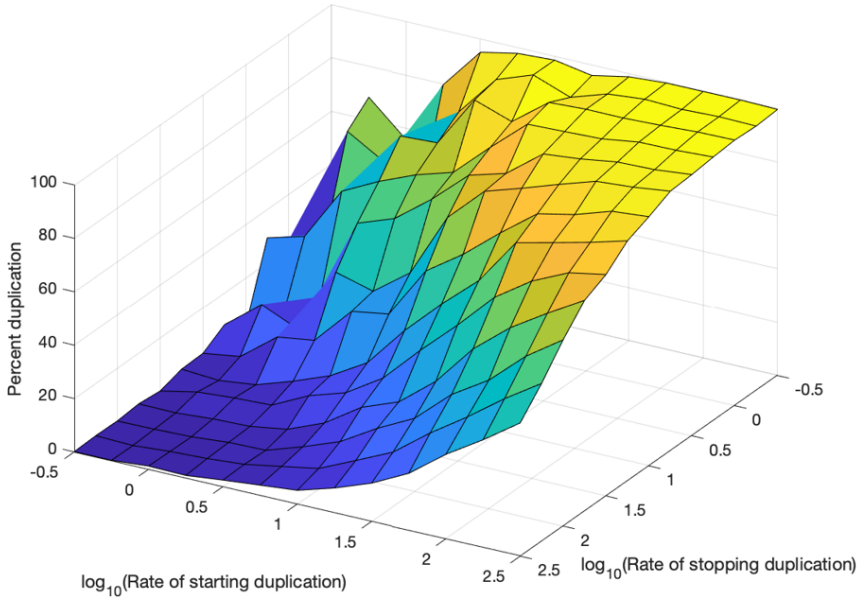

Figure 65: PD Based on Exponential Random Variable

4.4.7.3 Distribution Based Packet Duplication

We select exponential random variables to decide the rates of enabling and disabling the PD. We would turn the PD ON and OFF randomly based on our exponential rates [94]. If both BS and UE agree on the random number generator and seed, theoretically there would be no signalling overhead to turn on and off PD. This can also be termed as Zero signalling mechanism. An advantage of our method is that the UE and BS would be more aware of the event occurrences since random ON and OFF can always be studied and required resources can always be made available beforehand. As you can see in Fig. 65, the exponential random rate of starting duplication is plotted on the x-axis and exponential random rate of stopping duplication is plotted on the z-axis. The 3D plot

shows how the packet duplication is impacted due to the coordination on these two exponential random variables impacting the PD ON and OFF rates. A 100% PD is achieved when the log of the starting duplication rate is about 2.5 (rate equals $10^{2.5} = 316$ starts per second) and corresponding log of the stopping duplication rate is at -0.5 (rate equals $10^{-0.5} = 0.32$ stops per second). The reverse is for the 0% PD where the log of the starting duplication rate is about -0.5 and corresponding log of the stopping duplication rate is at 2.5.

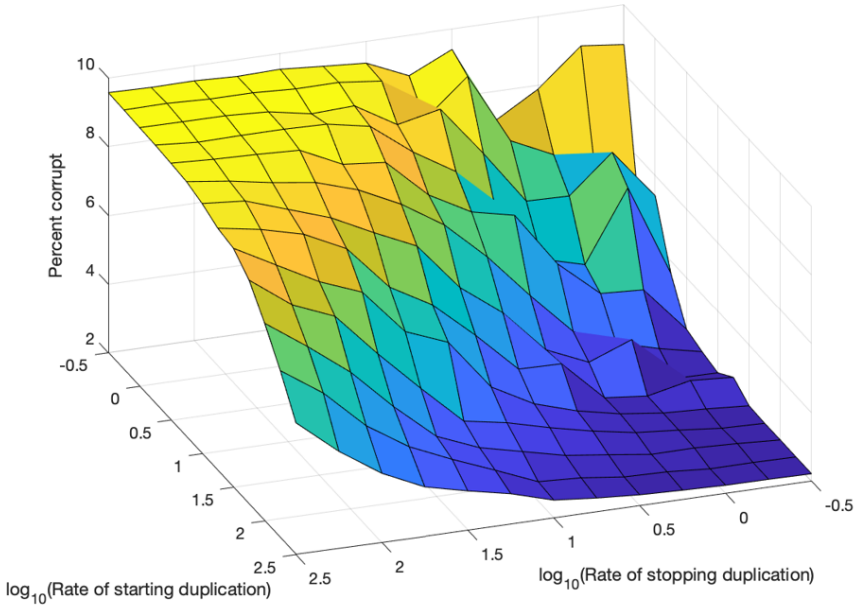

Figure 66: Corrupt PD Based on Exponential Random Variable

The corresponding corrupt packets are shows in Fig. 66. Similar to the above plot, a lower starting duplication rate along with a higher stopping duplication rate yields over 9% corrupt packets. And a maximum starting duplication rate with a lower stopping

duplication rate gives corrupt packet percentage close to 2%. Lastly, Fig. 67 shows the number of times PD is triggered in the entire communication. It can be clearly seen that when both the starting duplication and stopping duplication rates are high, the number of switches is also high. The PD switches are the lowest when both these rates are at their lowest. This study can be used to understand and analyze the RF channel and the actual amount of PD can be determined. In this study, we have used the same mean starting and stopping rates throughout the simulation. But as seen Fig. 50, the mean SINR values change over the course of a simulation when affected by buildings and distance. In [94], the best starting and stopping rates are based on average SINR. So it would be advantageous in practical application to have some adjustment of starting and stopping rates over time; however, these would change infrequently.

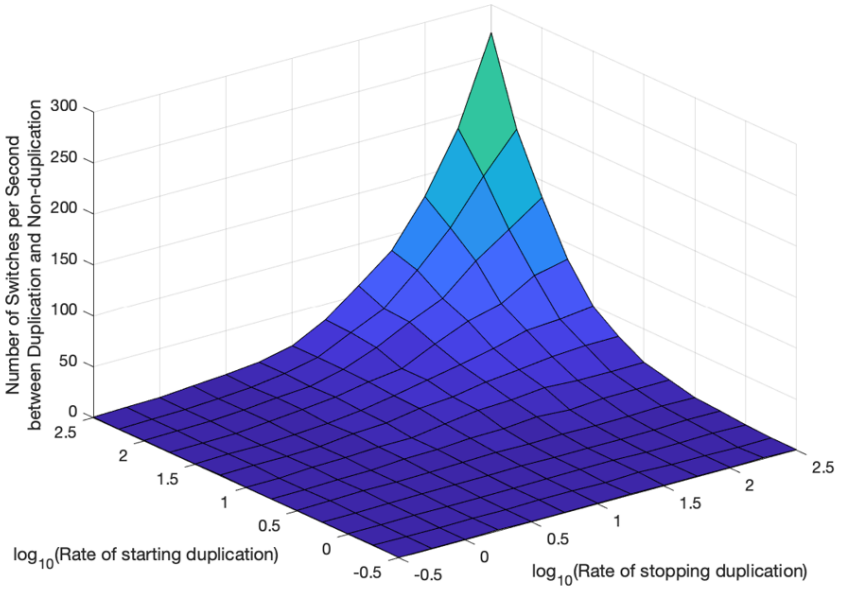

Figure 67: Actual Number of Times PD is Triggered

4.5 Conclusion

The handover process has to be efficient to meet requirements and not waste resources. FDOP provides a direct relationship with quality of connection. FDOP requires earlier handovers than SINR-based, but with more time to complete that handover. In some cases, usable coverage is less than expected, but the handover process can also be more relaxed in the time required for completion. A key result, however, is the use of fractional packet duplication along with FDOP. We can only use the number of duplicated packets that are really required. Since FDOP provides a direct understanding of user and application quality, the duplicated packets can be limited to only what is required. It also helps the network make an early decision and protects the user from losing any connection with the network, fulfilling the URLLC network requirements. Our work can be expanded to incorporate the Stochastic geometry of the surroundings to make FDOP more accurate.

The radio resources are very limited and need to be very efficiently used to meet the reliability and low latency requirements in 5G. Multi-Connectivity adds Spatial Diversity but also helps with beam forming and massive-MIMO in case of mmWave connections. Our proposed Adaptive Fractional Packet Duplication scheme will allow for the flexibility in the network to turn ON and OFF the PD. Our multiple schemes using the SINR or Fade threshold are the most effective ones as they require small changes in real network algorithms. Since a complete PD is wasteful over the entire transmission time, all our simulation results clearly show when and where PD will be effective and help understand the channel conditions even clearly. A network operator can thus decide on PD depending on the resource availability and application requirements. Future work can

include more than two connections and can also include the WiFi6 standards to improve data rates and help with cellular network offloading.

CHAPTER 5

DEEPSLICE AND SECURE5G: A DEEP LEARNING FRAMEWORK TOWARDS AN EFFICIENT, RELIABLE AND SECURE NETWORK SLICING IN 5G NETWORKS

5.1 Introduction

This chapter is based on two of our papers published on optimization of secure 5G network slicing, "Deepslice: A deep learning approach towards an efficient and reliable network slicing in 5G networks" by A. Thantharate, R. Paropkari, V. Walunj, and C. Beard, published in the 2019 IEEE 10th Annual Ubiquitous Computing, Electronics and Mobile Communication Conference and the second paper "Secure5G: A deep learning framework towards a secure network slicing in 5G and beyond," by A. Thantharate, R. Paropkari, V. Walunj, C. Beard, and P. Kankariya, published in the 2020 10th Annual Computing and Communication Workshop and Conference (CCWC). I want to thank my co-authors for their contribution towards this research and their help with writing the papers.

The evolution of the 5G network has opened an arena of possibilities and capabilities that are unavailable in the present day's 4G/LTE network. The critical aspect of 5G wireless digital transformation will enable the network operators to move their network functions into virtualized core and cloud, leading to new vertical use cases for businesses, enterprises, and consumers. The growing opportunities have resulted in a race among the

network operators and service providers to deploy network slicing functions. Network Slicing provides ease of operation and flexibility to create multiple logical networks on top of a commonly shared physical network infrastructure. Network Slicing will also allow for the orchestration of a dedicated end-to-end network for specific applications at scale while maintaining their respective service demands and needs.

We basically break this down into two parts, first we propose the DeepSlice model in section 5.3 and then we discuss our Secure5G model in section 5.4 with various use cases. The main goals of our DeepSlice model are (1) appropriate selection of a network slice for a device, (2) correct slice prediction and allocating enough resources to that slice based on the traffic prediction, and (3) adaptation of slice assignments in cases of network failures. The key tools for accomplishing these goals are deep learning neural networks. This work makes use of ML and Deep Learning Neural Networks (DLNN) to help make the most efficient and optimized selection of network slices for devices and/or services. Our DeepSlice model also analyzes the overall traffic pattern and can predict future traffic, so it can allocate resources, in advance, to the most appropriate slice.

For the Secure5G part we have addressed the issue of DDoS attack in 5G network from the UE perspective. We have developed a 'Secure5G' model based on deep learning techniques for Network Slicing function in 5G with an aim to (1) identify the incoming connection request and assign the most optimal slice based on the device type, (2) verification of the connection request if it is legit or a potential threat, and (3) implementation of an action, either assignment of appropriate network slice (valid request) or transfer to the quarantine slice (malicious request). The aim of the Secure5G is to mitigate the DDoS

initiation attacks by UE's, by making sure UE's can access network slices only after being authenticated and/or authorized minimizing the risk of denial of service. Secure5G protects the system in case of Volume-based flooding attack and in case where hackers mask the device identity and tries to exploit the network slice by requesting system access with low secured slice instance. The Secure5G model analyzes the overall traffic pattern and can predict future traffic so that it can allocate resources, in advance, to the most appropriate slice securely.

3GPP Rel. 15 has covered several technical specifications, important ones include the new security measurement for rogue, false base stations by masking the subscriber permanent identifier (SUPI), so that the rogue base station cannot track the subscriber in 5G network. 5G Globally Unique Temporary Identifier (5G-GUTI) is another improvement where UE and RAN have a mandatory requirement to refresh the GUTI from initial registration to mobility registration update. At the time of writing this work, 3GPP Rel. 16 which is expected to be available mid-2020 and list of work items those are being considered for standardization indicates the security requirements for Enhanced Network Slicing, URLLC for 5G Core and Cellular IoT [119].

The vision of next-generation 5G networks is to improve the capacity, coverage, security and connectivity of existing 4G networks. Network operators are designing the mmWave network to meet high capacity demand and relying on Sub 6 GHz 4G/LTE network for coverage. The current release of '5G NR' is based on 3GPP Release 15, which takes advantage of both sub-6 GHz and above 24 GHz to achieve substantial peak throughputs and low latencies. Current 5G deployment is on an overlay of the 4G LTE

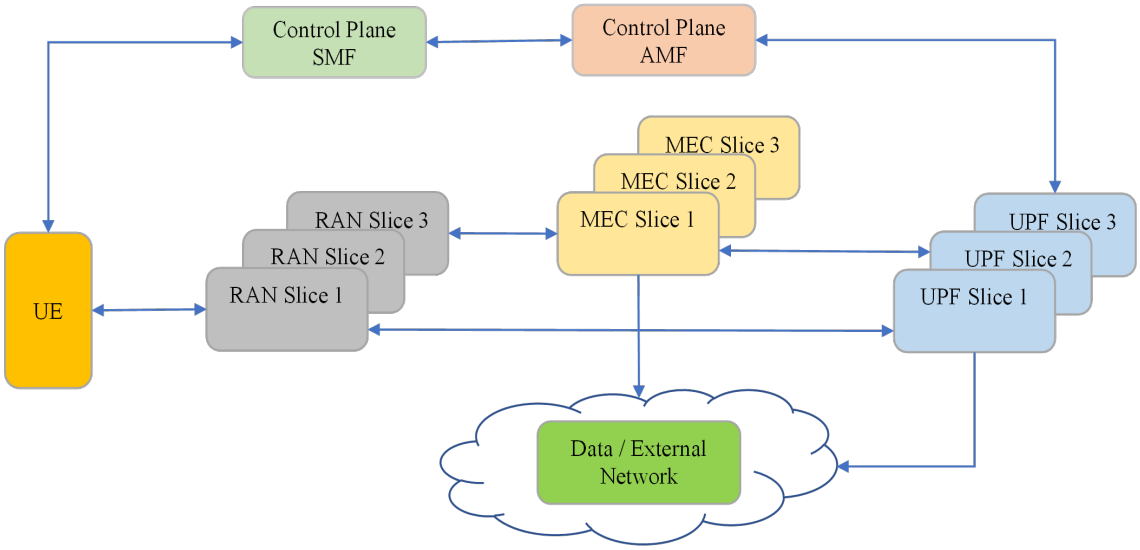

Figure 68: 3GPP based Network Slicing Framework

network and different service providers are following a different approach. Mobile Network Operators (MNO) are deploying or planning to deploy 5G using two architectures - Non-Standalone (NSA) and Standalone (SA). NSA is an evolutionary step for network operators to offer 5G services without building out a brand new dedicated 5G core network. In Non-Standalone, 5G-enabled smartphones will connect to 5G frequencies for data-throughput purposes but still use 4G/LTE for all control plane signaling to the cell towers and servers. On the other hand, Standalone will have its own dedicated core, and a UE will be able to use the 5G NR core for control plane signaling as well. SA will also support the growth of new cellular use cases such as Network Slicing, Control and User Plane Separation (CUPS), Virtualization, Multi-Gbps support, URLLC, and other such aspects that will be natively built into the 5G SA Packet Core architecture. We cannot ignore the fact that the threat and security loopholes with the evolution of these new

networks are growing substantially.

Fig. 68 illustrates the general layout of Network Slicing in 5G mobile networks. A core network slice is composed, on the control plane side, of a Session Management Function (SMF), which manages protocol data unit (PDU) sessions, and, on the user plane side, of a User Plane Function (UPF) and possibly other functions. Some network functions on the control plane are common and shared among multiple RAN and core network slices. An excellent example of a shared function is the Access and Mobility Management Function (AMF). To create a robust Network Slicing framework per 3rd Generation Partnership Project (3GPP) specifications [119], Network slices should be isolated from each other to avoid control plane congestion on one slice (e.g., using one SMF in slice dedicated for broadband applications) to affect the control plane of other slices (e.g., to affect potentially critical IoT applications). Since some common core network functions like AMF, PCF and UDM, etc. are shared between multiple dedicated core network slices, the interaction between shared NFs and NFs in dedicated network slices must be isolated from each other as well.

We introduce 5G network slicing and deep learning concepts in Section 5.1, most of the background work details are in Section 5.2, we explain our DeepSlice model in Section 5.3 with results and discuss its application for our use cases of slice prediction for unknown device types, load balancing and network failover scenario. In Section 5.4 we propose Secure5G model with use case evaluation and results. Finally, in Section 5.5 we conclude our work and propose possible future extension.

5.2 Related Work

Authors in [120] explore the multi-tenancy nature of the 5G network slicing by demonstrating how the capacity of a MVNO is affected by the number of users, transmit power. SDN and NFV-based 5G core network architecture is defined in [121]. Ping and Akihiro propose an application-specific mobile network deep learning architecture to apply application specific radio spectrum scheduling in the RAN [122]. Authors in [123] propose a framework to prioritize network traffic for smart cities using a priority management SDN approach. Taewhan started work early on network slicing and discusses standardization of network slicing, network slice selection, identifying slice-independent functions and then proposes an architecture for slicing and the RRC frame [124].

Other than this work, no other work to our knowledge considers the easily overlooked but difficult problem of deciding which devices and connections should be assigned to which network slices. And our work here is the first to use deep learning to address this problem, which will provide benefits of fast, flexible, accurate and informative decision making in the process. The authors in [27] contrasts Fade Duration Outage Probability (FDOP) based handover requirements with the traditional SINR based handovers methods in cellular systems. Another SDN and NFV based work on slicing demonstrates dynamic data rate allocation and the ability to provide hard service guarantees on 5G new radio air interfaces [125]. Many industry white papers and network surveys have been published and an Ericsson mobility report predicts the growth of mobile devices, 5G network connections and the overall data usage in coming years [3]. As for network intelligence, the authors in [22] represented handovers using matrix exponential distributions

for public safety and emergency communications, which helps make handover decisions more accurate considering all the different parameters involved in the decision process.

Authors in [126] present network survivability framework in 5G networks demonstrating network virtualization with multiple providers which necessitates network slicing in 5G. Virtualized networks or slices of virtualized networks are selected and assigned based on QCI and security requirements associated with a requested service in [127]. Campolo, et. al., share their vision about V2X network slicing by pin-pointing key requirements and providing a set of design guidelines, aligned with ongoing 3GPP standard specifications and network softwarization directions in [128]. The proposed model in [129] enables a cost-optimal deployment of network slices allowing a mobile network operator to efficiently allocate the underlying layer resources according to its users' requirements. However, none of their work considers the possibility of multiple service requirements requested by the same device, especially requested by an unknow device. Also, network slice load balancing and future prediction of traffic is unique in our work, especially with the use of ML and DL neural networks.

The 5G Network Slicing concept is fully utilized to manage the network traffic and route the connections to the most appropriate slice using DLNNs and understanding of what the connection demands in [24] A mathematical model that can provide on-demand slice isolation as well as guarantee end-to-end delay for 5G core network slices is proposed in [130] to proactively mitigate Distributed Denial-of-Service attacks in 5G core using slice isolation. The network slices relying solely on common infrastructure cannot meet highest isolation requirements and therefore authors in [131] introduce different

novel provisioning models for 3rd-party slices and discuss their isolation properties. Authors in [132] propose an efficient and secure service-oriented authentication framework supporting network slicing and fog computing for 5G-enabled IoT services. The also introduced a privacy preserving slice selection mechanism to preserve both configured slice types and accessing service types of users. [126] proposes a 5G network architecture framework with network virtualization among multiple providers, and a self-organizing ad hoc network among the eNBs that may use another provider for network resilience when the aggregation network and the backhaul network fail.

As for network intelligence, the authors in [22] [27] represented handovers using Markov Chain Matrix Exponential (ME) distributions for public safety and emergency communications, which helps make handover decisions more accurate considering all the different parameters involved in the decision process. Three different models are demonstrated in [133] using the CoAP and MQTT application protocol, which aims at providing efficient mechanisms and methods for over-the-air (OTA) delivery of software updates and security patches to IoT devices. Authors also evaluate which protocol is better suited for proposed models and applications. [134] applies a deep auto-encoded dense neural network algorithm for detecting intrusion or attacks in 5G and IoT network for flooding, impersonation and injection type of attacks.

5.3 5G Network Slicing, Machine Learning and Deep Learning

The current LTE architecture has a rigid framework that is not very flexible or scalable to adapt to diverse use cases. It often lacks customization when it comes to of-

fering any tailored business requirements or to meet specific business demands. With growing mobile data and consumer demands, business needs for faster connectivity and higher throughput cannot be fulfilled by today's 4G LTE network. Network slicing in 5G can cost-effectively deliver multiple logical networks over the same physical infrastructure. SDN and NFV together would allow us to manipulate these slices as and when needed without having to touch multiple different physical equipment in the network. Almost 'no-disruption' to any existing services is possible. Currently, service providers must configure and stitch together several components and equipment to achieve network slicing in 4G. Use of Access Point Name (APN) or Public Land Mobile Network (PLMN ID) are examples that service providers implement today for Mobile Virtual Network Operator (MVNOs), enterprise customers, etc. There is a lot of work done on optimization and efficient scheduling of radio and network resources; however, application or service-based resource allocation is a necessity and a must-have feature in 5G networks.

Operators have a huge amount of data traffic coming through their network which will increase with growing number of devices and additional services of 5G networks. This traffic can be segmented and dealt with individually and independently. It will benefit any service provider as they can now charge differently for each sliced segment and even adjust the cost for each slice, leading to a balance between business profitability and customer satisfaction. In addition, 5G network slicing allows service providers to build for all current use cases that have been around for a while, and some emerging applications and services as well. It will provide a 'one size fits all' approach. Each network slice can be isolated, have individual control and policy management systems. The inclusion of ML

here will allow us to analyze any unknowns and take necessary corrective actions. ML will provide network analysis of the huge data which can be studied further to efficiently and cost effectively modify any given slice as needed. DL for instance, as represented in Fig. 69, can trigger automation in the network to modify available resources and make changes on the go. DL will be responsible not only to provide and process, but also make an intelligent decision for network resource adaptation without any human intervention. It will also combine a variety of factors to make the best decisions, possibly too many factors for a human to consider at once or even be able to process in a short time.

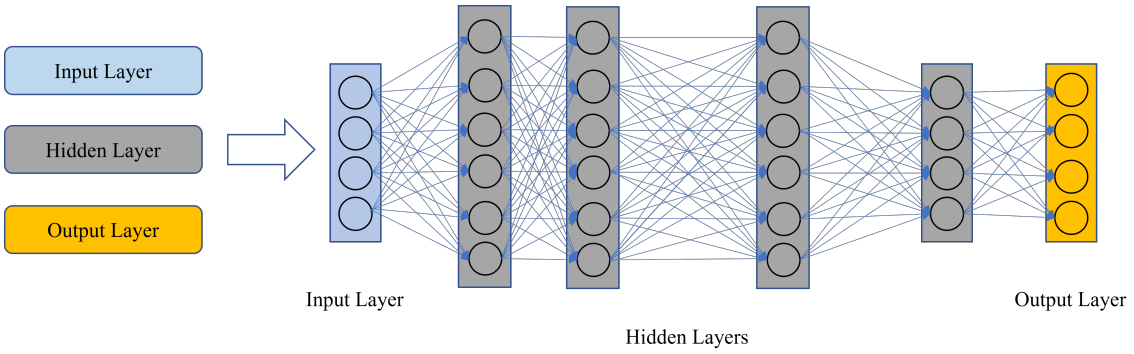

Figure 69: General Deep Learning Neural Network

DL will perform real-time analysis for any given slice to determine the network performance, create a potential baseline for performance, be proactive in anticipating problems, inspect different network elements, and find out if anything is abnormal. A simple example could be on a slice for fixed wireless enterprise network, wherein if the network sees a sudden demand increase, automation can add more capacity in real time to provide efficient communication. This will help to create any newly required services or slices in the network. Automation will facilitate all this in a shorter timespan without caus-

ing any performance issues to an on-going session. Current hurdles in implementation of network slicing are organizational, as one will have to touch several different pieces of hardware and groups in a service provider network, to make a single change. The programmability capabilities of 5G will provide flexibility to seamlessly stitch together an end-to-end service for any application. A typical consumer would request parameters like data rate, latency, mobility, isolation, power constraints, etc. Accordingly, a specific network slice type is provisioned if the existing network slice instance does not have enough capacity and associated network functions are initiated on demand.

Each use case receives an optimized set of resources in the network topology covering several SLA specified factors like connectivity, latency, priority, service availability, speed, capacity, etc. that suit the need of an application. The key parameters that are determined for network slicing are the slice type, bandwidth, throughput, latency, equipment type, mobility, reliability, isolation, power, etc. 5G enables enormous amounts of data collection, and this leads to the need of ML for big data analytics. Some of the most relevant and useful ML-based applications in the wireless industry are identifying and restarting sleeping cellular cells, optimizing mobile tower operations, faster wireless channel adoption, facilitating targeted marketing, autonomous decision making in IoT networks, real-time data analysis, predictive maintenance, customer churn, sentiment analysis by social networking, fraud detection, e-commerce, etc. ML implementation in Uber-like applications will have many advantages since Uber follows differential pricing in real time based on the demand, cars available, weather conditions, rush hour, etc. and so ML-based platform will allow for better accuracy and future prediction based on

enormous data from the past and in the present.

5.3.1 Proposed System Model - DeepSlice

Neural networks are widely used in the industry today, and their usage will only grow as the ever-growing devices on 5G networks generate massive data. Accurate analysis and decision making will be overwhelming for any human being and faster processing times are required. We first create an ML model and later build a DLNN to help decide which network slice to use for given input information. The developed 'DeepSlice' is then used to manage network load, slice failure conditions and detect the most appropriate slice for any new unknown device type connecting to the network. A statistical ML model is based on the Random Forest (RF) algorithm, and the DeepSlice uses a convolutional neural network (CNN) classifier. Both RF and CNN are widely used models in their respective domains. We use the exact same dataset for both our ML and DLNN models consisting of over 65,000 unique input combinations.

Our dataset includes most relevant KPIs from both the network and the devices, including the type of device used to connect (Smartphone, IoT device, URLLC device, etc.), User Equipment (UE) category, QoS Class Identifier (QCI), packet delay budget, maximum packet loss, time and day of the week, etc. These KPIs can be captured from control packets between the UE and network. Since our model will run internally on the network, all this information is readily available. We have multiple different types of input devices requesting access to our system. As shown in Fig. 70, these include smartphones, general IoT devices, AR-VR devices, Industry 4.0 traffic, e911 or public safety communication, healthcare, smart city or smart homes traffic, etc. or even an unknown device

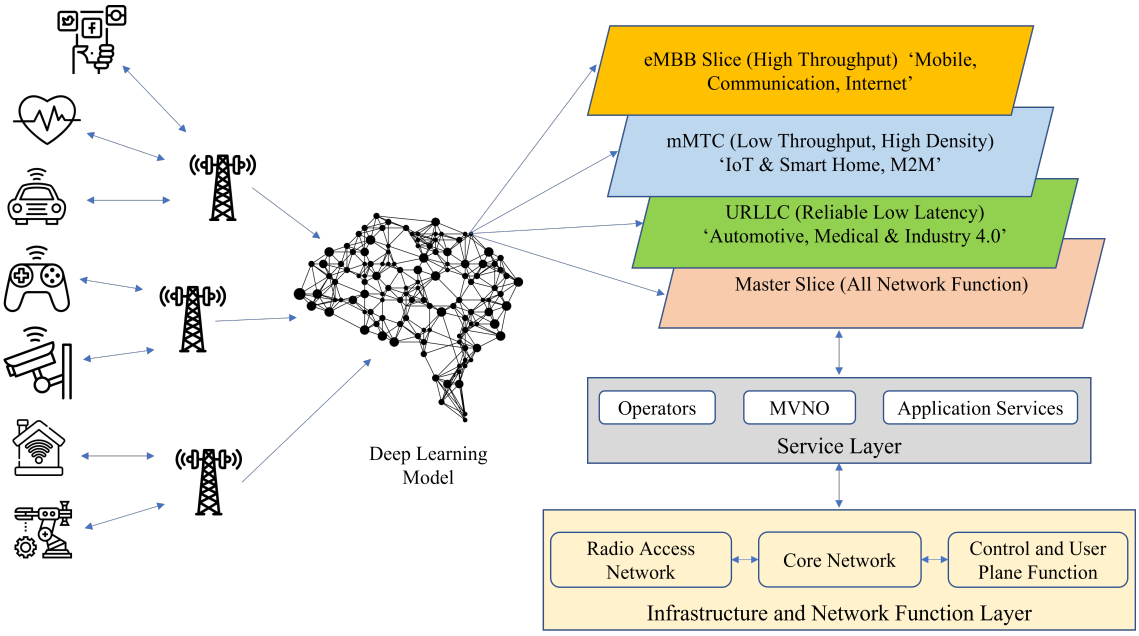

Figure 70: General Representation of our Deep Learning Neural Network Model 'DeepSlice' consisting of Network Slices

requesting access to one or multiple services. These have UE category values defined to them and the network also allocates a pre-defined QCI value to each service request. In 5G, the packet delay budget and the packet loss rate are an integral part of the 5QI (5G QoS Identifier), and we have them included in our model. DeepSlice will also observe what time and day of the week is the request received in the system. All this information will be recorded and used by our DLNN to make smart decisions in the present and efficiently predict network resource reservation for the future.

In , we have shown highlights of the features of our simulation model. The second column shows the average time duration spent in the system by each of the incoming requests. All these incoming requests are directed to one or more of the network slices as predicted. We have also considered some variations in the traffic types; mMTC de-

vices can be further categorized as ones requiring a continuous connection link and others needing only a momentary connection to send data periodically. Smartphone devices can be used by common users to make phone calls, browse the web and at the same time by first responders in an emergency (lower packet loss and packet delay). Our pre-defined slice categories include enhanced Mobile Broad Band (eMBB), Ultra Reliable Low Latency Communication (URLLC), massive Machine Type Communication (mMTC) and the Master slice. The Master slice is the slice that will have network functions belonging to each of the other slices. It can always act as a back-up slice, in a hot-standby, and will be used depending on the load on other slices.

In our proposed model, we predict the network load on each network slice based on the previous information of incoming connections and keep track of which output 'network slice' is being utilized the most. We then allocate incoming traffic to the network by efficiently distributing them between all the slices as desired. We have used Keras which is a deep learning library in Python for our model simulations. A DLNN is required as there are no clear sets of rules for how each incoming device type should be treated. Cellular handovers, for example, are based upon several network factors. With

Table 4: Feature highlights of our DeepSlice simulation model

Input Type	Duration	Packet Loss Rate	Packet Delay Budget (ms)	**Predicted Slice**
Smartphone	300 or IoT	$10^{-2}/10^{-3}/10^{-6}$	60/75/100/150/300	eMBB/mMTC
IoT Device	60	10^{-2}	50/300	mMTC
Smart Transportation	60	10^{-6}	10	URLLC
Industry 4.0	180	$10^{-3}/10^{-6}$	10/50	mMTC/URLLC
AR/VR/Gaming5	600	10^{-3}	10/50	eMBB
Healthcare	180	10^{-6}	10	URLLC
Public Safety / E911	300	10^{-6}	10	URLLC
Smart City / Home	120	10^{-2}	50/300	mMTC
Unknown Device Type / Home	60/120/180/300	$10^{-2}/10^{-3}/10^{-6}$	10/50/60/75/100/150/300	eMBB/mMTC/URLLC

every new scenario, an intelligent network can learn and adapt very quickly to changes or new requirements compared to traditional algorithms. DLNN can help identify and accommodate the unknows in the network.

5.3.1.1 Machine Learning with Random Forest Algorithm

When we have a well-structured data with multiple attributes, use of Random Forest (RF) along with DLNN is the most recommended option. RF is a supervised learning model and mainly used to build predictive models for both classification and regression problems. The main reason for selecting RF for our model over k-Nearest Neighbor, Naive Bayes, or Decision Tree is simply because of the nature and amount of data we have in our dataset. We have around 65K unique inputs, and all this data is well structured, so RF reduces the risk of overfitting by using multiple sub-trees. RF is useful to quickly classify input data into any pre-defined category. RF runs efficiently on a large database and produces accurate predictions. Most importantly, it estimates any missing data and maintains the accuracy even when some input data is missing.

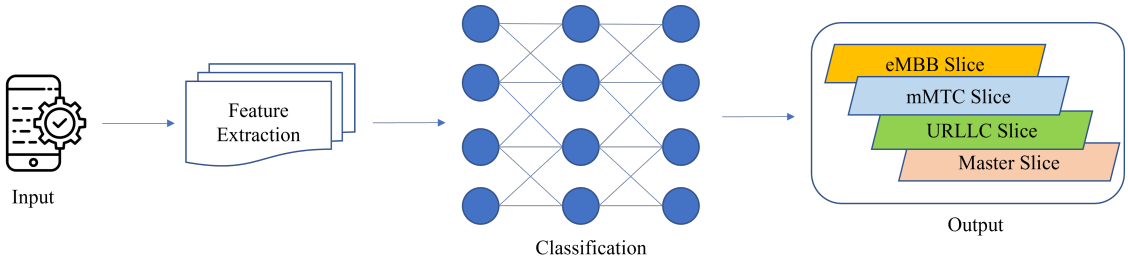

Figure 71: Machine Learning Model

Fig. 71 and Fig. 72 illustrate the typical ML modeling with decision trees and predicting the output with majority voting. As per our input dataset, we have about 8

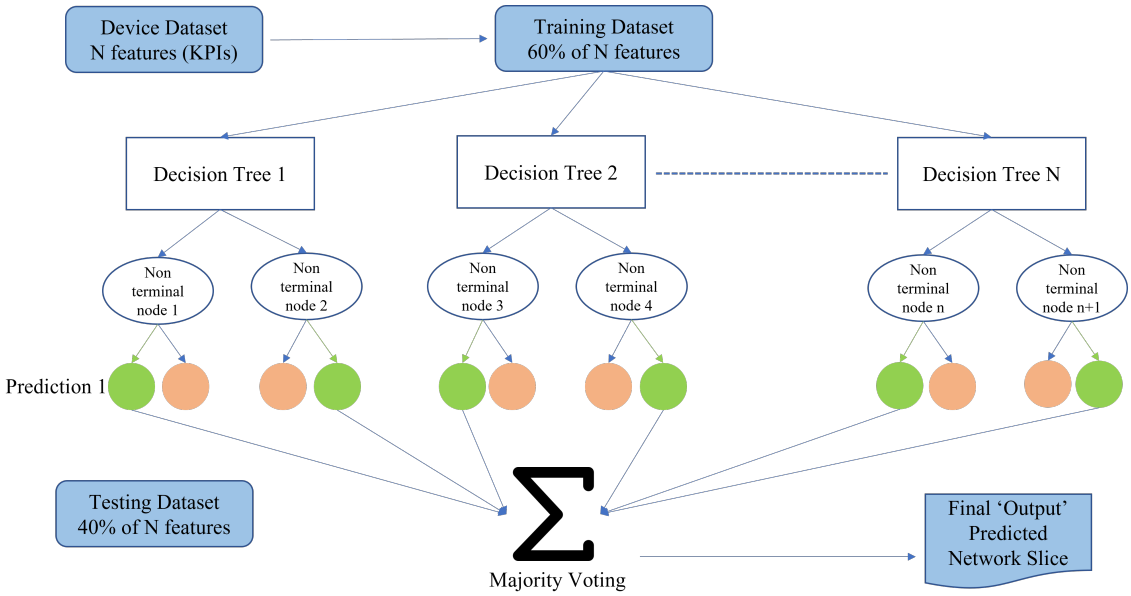

Figure 72: Random Forest Decision Tree based ML Model

different input strings that will together contribute towards a decision that the model will make. And it can very well happen in the real-world scenario that one or more among the 8 inputs may not be received and our model still must predict an output. During training of our data, RF constructs multiple decision trees based on inputs, each branch of a tree represents a possible occurrence or response. We use 70% of our input dataset to train our model and the remaining 30% was used for predicting the classifier accuracy. The RF algorithm in our ML model gives high accuracy.

5.3.1.2 Deep Learning Neural Network

The DLNN works best when the data is unstructured and huge. We use the same dataset to train multiple neurons of our DLNN, and it predicts the correct network slice based on any input from the UE information. Our DLNN can predict very accurately and

we utilize this functionality to select the correct slice for unknown device types. It helps redirect traffic to the Master slice if load balancing is required in the network slices, and in case of any slice failure in the network. In our proposed DeepSlice model in Fig. 73, we predict the network load of each network slice based on the incoming connection and keep track of which output 'network slice' is being utilized most. We then allocate incoming devices to slices by efficiently distributing between the eMBB, URLLC, mMTC or the master slice depending on the load and the output predicted by our model.

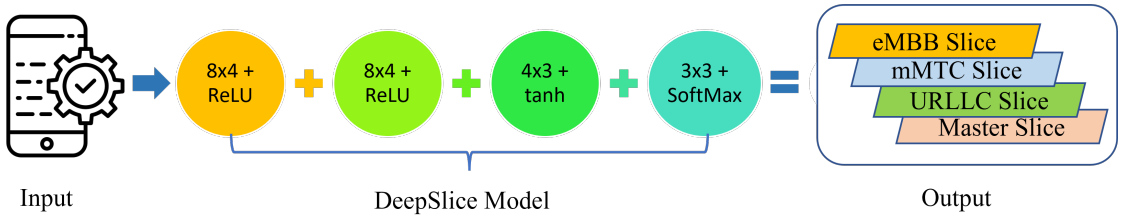

Figure 73: 'DeepSlice' DLNN Model Overview

Approximately a quarter million user connection requests were generated in a 24-hour simulation of which 40% was eMBB, 25% mMTC and 35% URLLC. Fig. 74 above shows the simulated DLNN model run for 24 hours giving the number of users being served at an instance. The plot begins when our model had reached a steady state which was at the 1-hour mark. Based on the Table 4 information, all incoming traffic has a pre-defined time-to-live (TTL) and so only a fraction remains alive every second. For example, eMBB active user average count was 275 at any given instance. URLLC and mMTC users were allotted short TTL compared to the eMBB which is why we have more users alive for broadband services. This can help analyze the user pattern and will

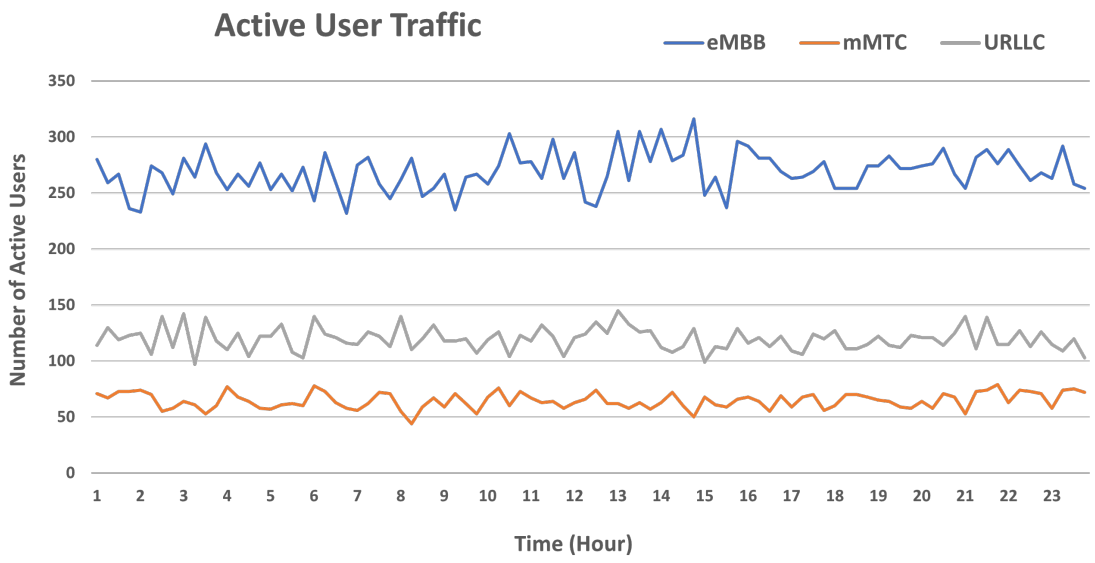

Figure 74: Active User Count in the Network observed every 15 Minutes

allow for automated decisions based on the retrieved input information from the connected device.

DeepSlice will eventually learn and understand what kind of a device goes to what slice and it will evolve over time to be able to predict future connections requiring a specific service or a network slice. It can help prepare the network for any new connections by properly allocating resources in advance; this will save any delays later. Our dataset includes day and time of any connection which can also help the network predict number of connections in the future at any given time and would be aware what network slices would be required or requested by those connections based on learning from the past information.

5.3.2 Use Cases and Performance Evalucation

In this section, we evaluate DeepSlice and verify how it can be used to provide slice prediction, load balancing and network availability. In our first use case, we validate our approach by demonstrating how slices are accurately selected for any unknown device types requesting connections to the system. Our second use case of load balancing involves efficient utilization of each of the available network slices. If any individual slice utilization exceeds a certain threshold of its total available resources, our model will direct any new connections to the master slice that is otherwise required to carry the device when a slice utilization exceeds a pre-defined threshold. Our third use case depicts a slice failure scenario where all that traffic will route to the master slice instead and prevent any loss of service during failure of the slice. DeepSlice will capture the time of any connection failure and some attributes around the failure; the next time it can try to isolate the issue and be prepared in advance.

5.3.2.1 Unknown Device Type

DeepSlice model is trained using our dataset of multiple unique inputs based on network and device KPIs. Our cross-validation accuracy was over 90.62% (Fig. 79) which included the entire test dataset of new input scenarios, those not used while training. We also included certain unknown device types with randomly selected parameters. Slice prediction accuracy was 95% for unknown devices. Table 5 shows a few unknowns and how only a portion of input information was used to correctly determine the network slice to be used.

Our training dataset included 6 to 8 parameters in every input, but our model requires a minimum of 2 or 3 input KPIs, to determine the services requested and allocate the correct slice. This is very essential, since a lot of devices with various capabilities request different services at different times. An industry 4.0 IoT application requires very low latency in pharmaceutical environments (URLLC), whereas the same type could also be used for monitoring production lines, which would require periodic connection and very low throughput (mMTC).

5.3.2.2 Load Balancing Scenario

We use the same DLNN but assume that one slice would be overutilized if the number of connections exceed a threshold, say 90% usage in our case. Fig. 75 shows an eMBB slice is detected to have over 90% utilization with its traffic to go over the set threshold, so, the master slice acts as backup for any new eMBB connections. Our DeepSlice can realize this overload and can be prepared next time to redirect traffic without causing one specific slice to be overloaded. When compared with Fig. 74, the master slice takes over the excess traffic as shown in Fig. 76.

Table 5: Slice prediction for unknown device types

Input Type	Technology	Packet Loss Rate	Packet Delay Budget (ms)	**Predicted Slice**
Unknown Type - 1	LTE/5G or IOT or IoT	10^{-3}	50	eMBB/mMTC
Unknown Type - 2	IOT	10^{-2}	50	mMTC
Unknown Type - 3	IOT	10^{-6}	10	URLLC
Unknown Type - 4	IOT	10^{-2}	300	mMTC
Unknown Type - 5	LTE/5G	10^{-2}	100	eMBB
Unknown Type - 6	LTE/5G	10^{-6}	100	eMBB

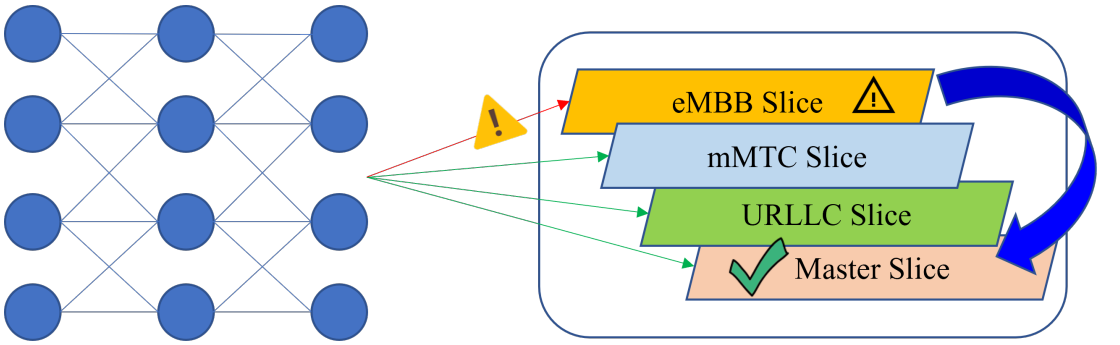

Figure 75: Slice Utilization exceeding a pre-defined Threshold

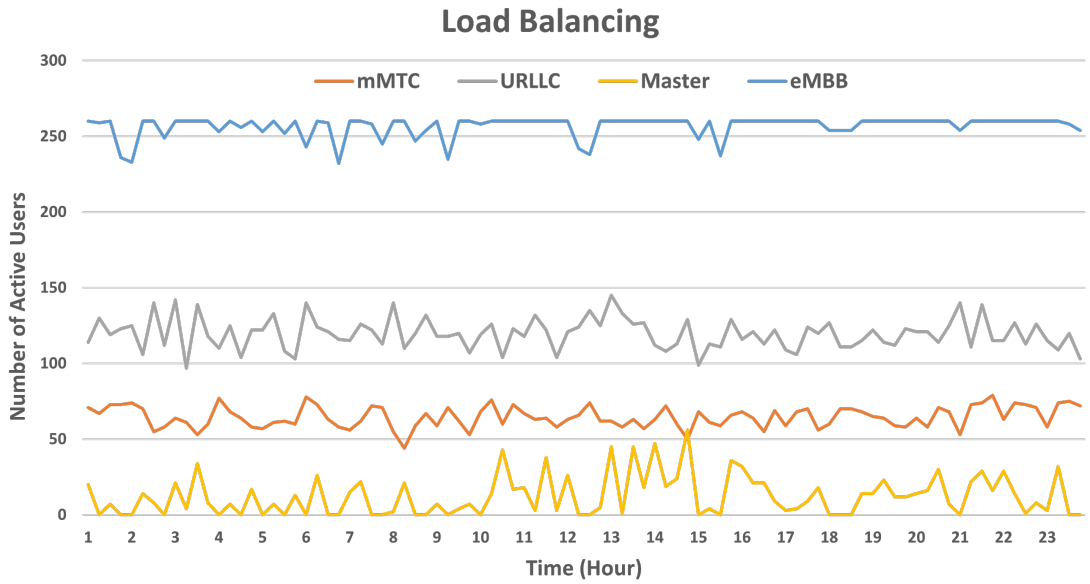

Figure 76: Slice Utilization exceeding a pre-defined Threshold

5.3.2.3 Network Slice Failure Scenario

In this case, we assume a complete failure of a specific slice, specifically eMBB as shown in Fig. 77. Now the DeepSlice will direct all new eMBB related traffic to the

master slice and avoid any loss of traffic transmission in the network. However, any ongoing communication on that slice would be impacted and all existing connections are lost due to sudden slice failure. This is recorded by the system, say for example, date and time, and care will be taken next time to avoid loss of all ongoing connections. Fig. 78 shows that our simulated model had failures on the mMTC slice for a period of two hours from 3hr to 5hr and on the eMBB slice for another two-hour period 16hr to 18hr. The master slice was identified as a backup and used to redirect this traffic during those slice failures. We had substantial resources reserved in the master slice for each of our network slices in terms of capacity and processing speed.

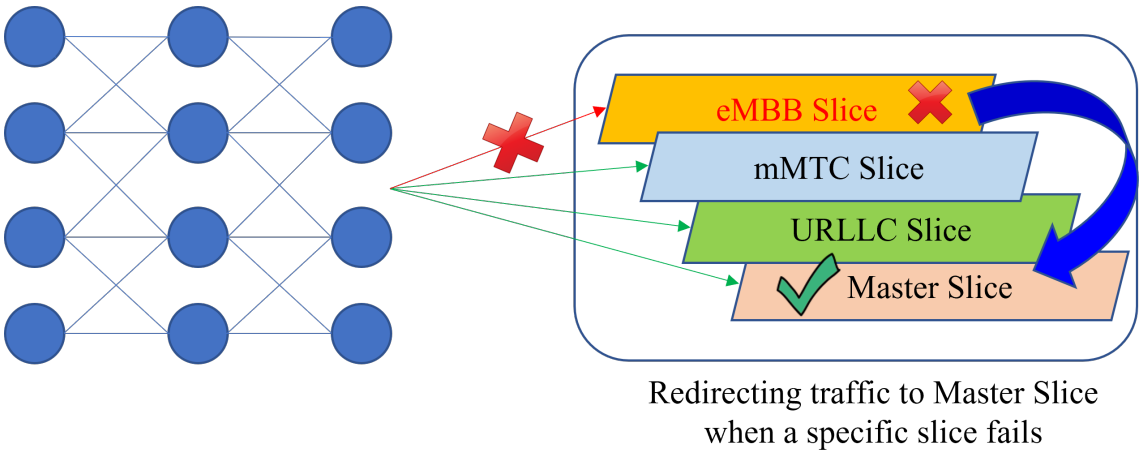

Figure 77: Network Slice Failure and re-direction to Master Slice

We have run our simulation for a period of a day and later for a whole one-week period to get close to real-time results. The randomly distributed average connections received in an hour did not change between a day and a week. One-week simulation produced almost two million service requests. We also used multiple unknown device types and our model was able to maintain the accuracy for prediction of slices. Figs. 7a

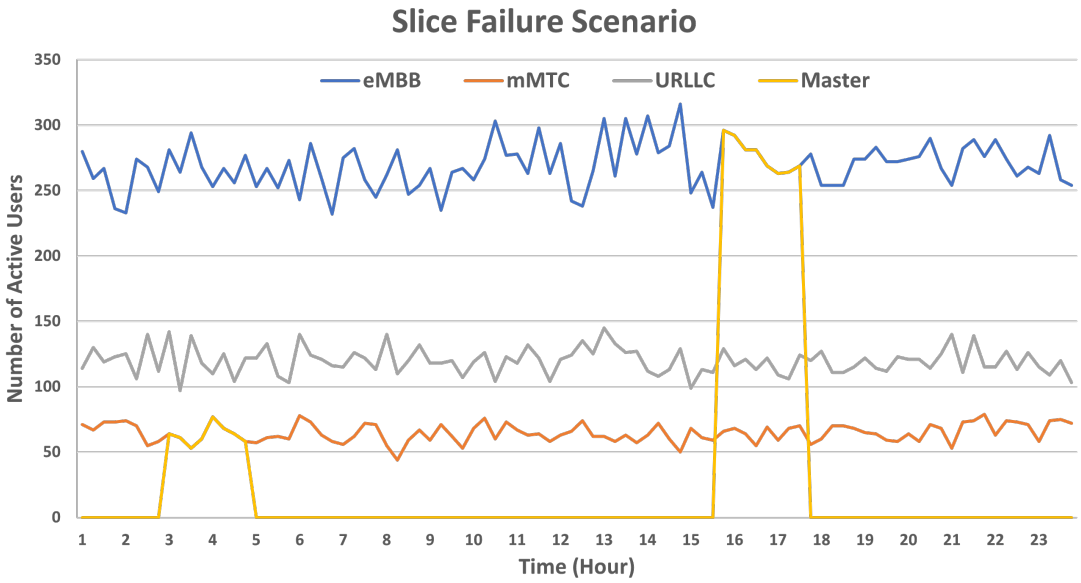

Figure 78: Network Slice Failure and re-direction to Master Slice

and 7b shows the accuracy or measure prediction quality of our model.

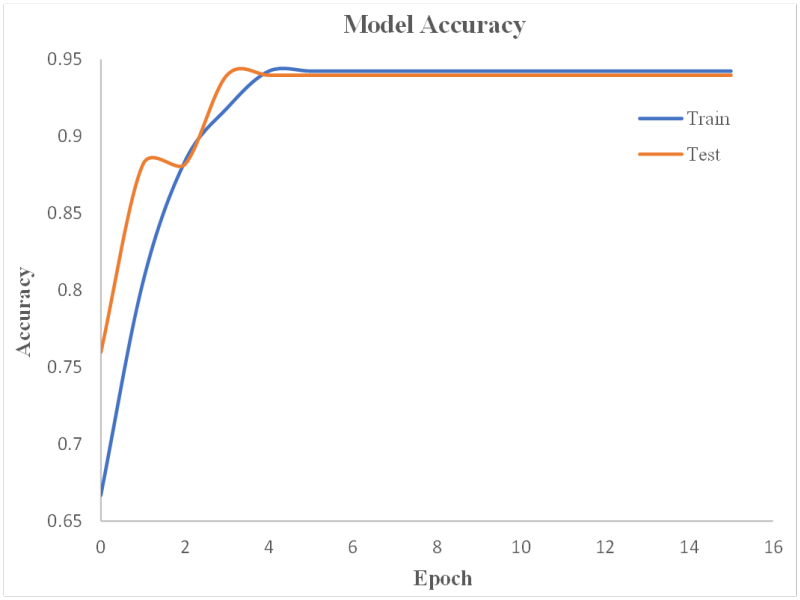

Figure 79: Training and Validation Accuracy

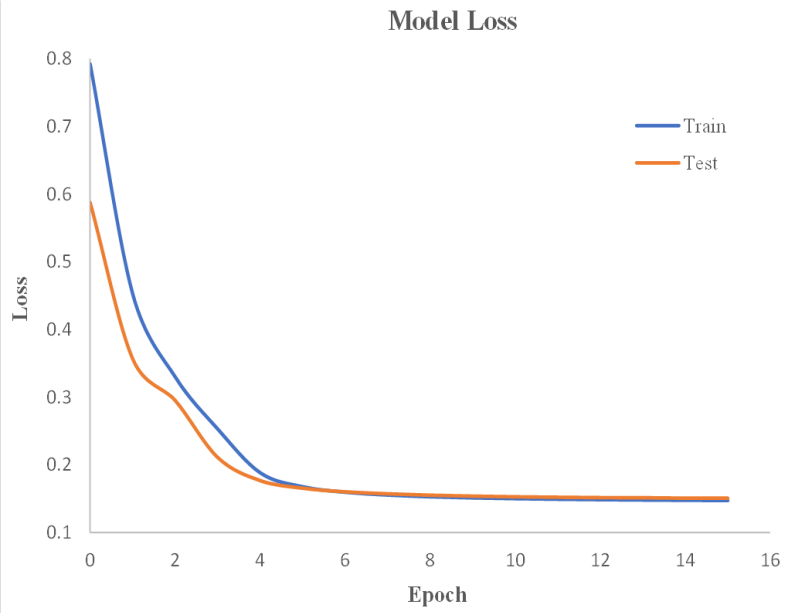

Figure 80: Training and Validation Loss

5.4 5G Security, DeepSlice and Secure5G

In general, the threats and security challenges faced by the 5G ecosystem are the same as encountered by 4G/LTE today. 5G networks, in addition, will have specific requirements on throughput, latency, and security to meet the service level agreements for diverse applications and services, especially with a diverse ecosystem for IoT devices. F-secure threat report 2019 [135] indicates that 99.9% of the attack traffic comes from bots or some automation tool, IoT bot activity represented 78% of the malware network activity (detection events) across carrier networks; more than triple the rate seen in 2016, since the introduction of biggest DDoS attack in history with Mirai botnet. Mirai was a brute force password guessing attack on open telnet and SSH ports by scanning internet for open Telnet ports, then attempted to log in default passwords.

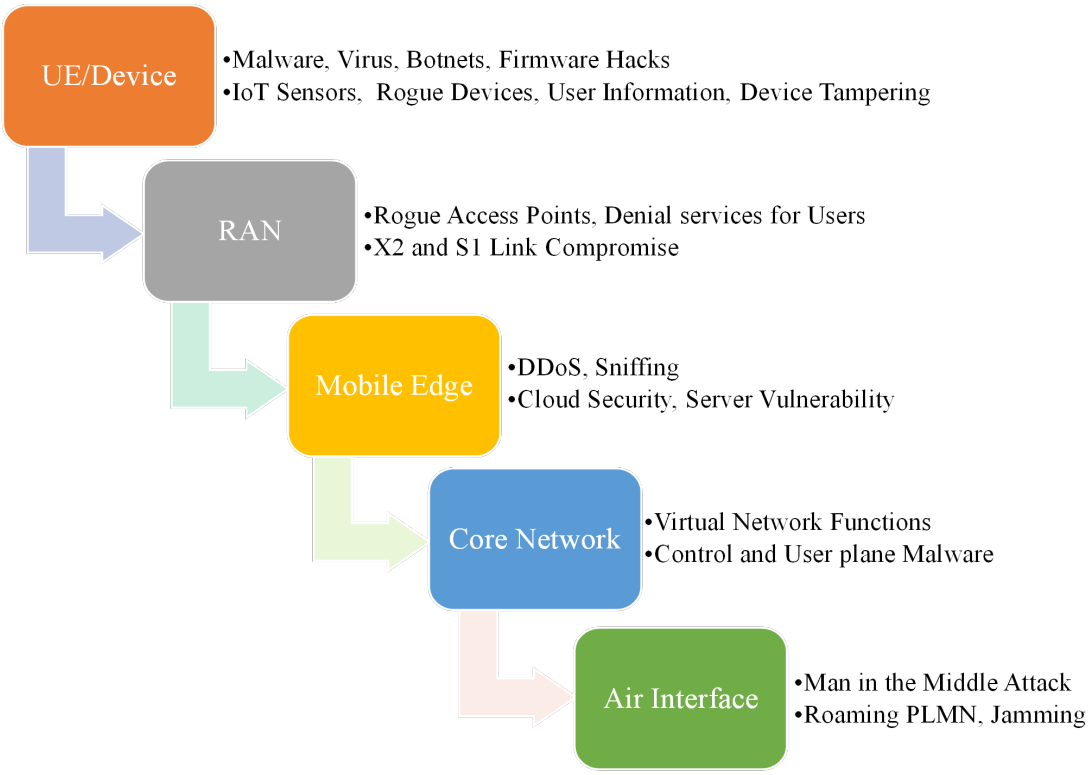

Figure 81: Common 5G Threats Vectors across Device and Network

Fig. 81 shows the various threat vectors that we classify in the 5G Network today. In typical DDoS attack, hacker floods the system by sending huge amount of bad traffic, false ping and connection request to the targeted network making bandwidth and resources unavailable or busy for normal traffic. UDP attack for example floods random ports on a remote host making host server busy and unresponsive, ICMP packets attacks ping packets continuously without waiting for replies causing system to clog, SYN flood exploits 3-way handshake mechanism and does not close the connection after receiving acknowledgement from the server causing server bottleneck situation. Ping of death which sends malicious pings with IP packets more than 65,535 bytes in length is another

form of cyber-attack with the intention to saturate and overwhelm the website or server and make resources unavailable for normal traffic.

Other network attacks include the International mobile subscriber identity (IMSI) catcher, where IMSI of a device is sent unencrypted over the radio and eavesdropped by hackers. The attacks on user plane and control plane on the core network and radio interface is also a common exploitation point for malicious actor and could lead to data injection or modification such as the Man-in-the-Middle (MitM) attack. Radio Access Network (RAN) attacks include the UE location tracking, malicious message insertions during the initial UE attach procedures. Bring your own device (BYOD) concepts are another threat concern for enterprise solutions, increasing floodgates, and data leakage opportunities for attackers. With less control over BYOD devices, systems are more vulnerable to attacks and device tampering. Battery life is a key aspect of the 5G, LTE-M and NB-IoT technology aims to power IoT device for battery life span up to 30 years, by disabling the power saving abilities of these IoT devices through injection of malicious code during initial attach could drain battery drain five to ten times faster than expected life.

Secure5G is an extension to the DeepSlice research work [24]. DeepSlice is a 4-layered deep learning neural network model comprising of input layer, output layer along with 2 hidden layers which was designed to predict the best optimal slice based on the diverse device input types requesting connection to the network with different Quality of service (QoS) Class Identifier (QCI) as shown in Fig. 73and Table 4. The incoming QCI request from user's device is used as the input parameters that are fed to

the DeepSlice model to identify the best slice for the requested service and allocate the network resources based on 65K unique dataset [136].

In Table 4, we have shown highlights of the features of our DeepSlice simulation model. The second column shows the normalized time duration spent in the system by each of the incoming requests. All these incoming requests are directed to one or more of the network slices as predicted. Smartphone devices can be used by common users to make phone calls, browse the web and at the same time by first responders in an emergency (lower packet loss and packet delay). Our pre-defined slice categories include enhanced Mobile Broad Band (eMBB), Ultra Reliable Low Latency Communication (URLLC), massive Machine Type Communication (mMTC) and the Master slice. The Master slice is the slice that will have network functions belonging to each of the other slices and acts as a back-up slice, in a hot-standby in case of any failures.

The Secure5G model primarily consists of User Equipment (UE) requesting service to the network for slice and resource allocation. In our model, we considered diverse input types and applications like smartphones, health devices, autonomous vehicles, AR/VR gaming, smart homes and cities, and Industry 4.0. We also included the Malware

Table 6: Feature highlights of our *'Secure5G'* and *'DeepSlice'* simulation model

Input Type	Duration	Packet Loss Rate	Packet Delay Budget (ms)	**Predicted Slice**
Smartphone	300 or IoT	$10^{-2}/10^{-3}/10^{-6}$	60/75/100/150/300	eMBB/mMTC
IoT Device	60	10^{-2}	50/300	mMTC
Smart Transportation	60	10^{-6}	10	URLLC
Industry 4.0	180	$10^{-3}/10^{-6}$	10/50	mMTC/URLLC
AR/VR/Gaming5	600	10^{-3}	10/50	eMBB
Healthcare	180	10^{-6}	10	URLLC
Public Safety / E911	300	10^{-6}	10	URLLC
Smart City / Home	120	10^{-2}	50/300	mMTC
Unknown Device Type / Home	60/120/180/300	$10^{-2}/10^{-3}/10^{-6}$	10/50/60/75/100/150/300	eMBB/mMTC/URLLC

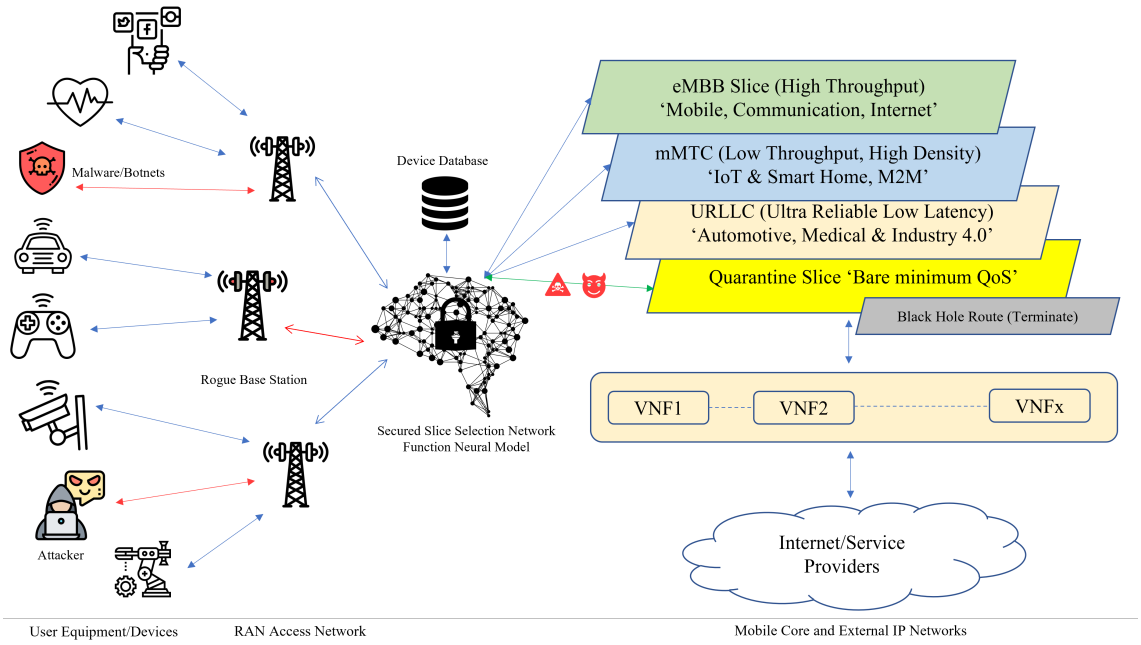

Figure 82: 'Secure5G' Secured Network Slicing Model Overview

botnets and hackers, as shown in Fig. 82. We introduced a concept of a 'Quarantine slice' along with the eMBB, mMTC, and URLLC standard slices. Quarantine slice has the network functions for a bare minimum QoS which allows a device to communicate to network with a very restricting set of requirements in case of slice attack, slice failures with bare minimum service to serve the user instead abruptly terminating the connection. Black hole routing concept has been used to terminate the connection permanently after observing the repeated malicious traffic pattern of the device(s).

5.4.1 'Secure5G' Model - Use Cases and Evaluation

In this section, we evaluate Secure5G deep learning model on how it can be used to proactively prevent DDoS attacks on a 5G network based on the incoming network

connections before it even reaches to the core network. Additionally, we are also observing the traffic pattern and QoS characteristics during slice allocation to detect anomalies. For instance, a device in an IoT slice whose traffic no longer matches IoT traffic patterns might trigger a warning for a potential attack. Secure5G knows if UE is accessing the unauthorized network slicing or requesting unauthorized operation, for example, changing QoS values for prioritizing eMBB over mMTC as an example. The model can also detect abnormal behavior by the subscribed user, for example, requesting access to multiple slices simultaneously repeatedly compared to their previous usage or usage in general from similar devices on that slice.

For evaluation, we considered two scenarios: Volume Based Attack (Flooding) and Masking Botnets. Secure5G has the capability to observe and learn the device request patterns and assign best optimal slice and model will evolve over time and be able to predict future traffic patterns and can be used for capacity forecast. The model also helps prepare the network for assigning slices to unknown (or new) application or service requests which are not known to the network and secure the system in case of slice DDoS attack or Slice failure as a result.

5.4.1.1 Volume based Attack

The volume-based attack is one of the widespread forms of cyber-attacks. Hackers disrupt the flow of normal service, typically, by flooding the target with a high volume of packets or connection requests, overwhelming networking equipment, servers, or bandwidth resources. We evaluated our Secure5G model by simulating the attack in a way that device(s) are making multiple connection requests to the network at the same time, and

our model is detecting the bad malicious traffic and blocking before it even reaches to the core network.

Per 3GPP specifications, a device can access multiple Network Slices simultaneously, slices can have a diverse configuration, which increases the possibility of security loopholes between slices. For example, if UE exchanges a sensitive date in one slice (enterprise) and publishes data on another slice (consumer), then there is a possibility of data leak between slices. The administration of data exchange between UE and different slices is a high level of security concerns, and this impact needs to be studied further. Security policies need to be defined either on the RAN or Core network slicing as the UE has no notion or control over which slice to request connection. The network operator should ask UE to re-authenticate for every network slice separately to check and validate if UE is meeting the SLA for every slice or not. Otherwise, a malicious UE can authenticate to a lower level security slice and get access to other slices through common network function and resource sharing.

As shown in Fig. 83, malware botnet and an attacker are trying to flood the network by sending multiple requests. If such attackers are allowed into the network, the system may run out of its capacity and crash; unable to handle huge traffic and might become unresponsive to new requests. A smartphone can only make one control Radio Resource Control (RRC) signaling connection request to the network during the initial attach and that too for a single network slice instance. However, if it makes multiple requests to multiple slices simultaneously, then such an unusual behavior will be identified as suspicious and Secure5G model will detect this anomaly and quarantine these incom-

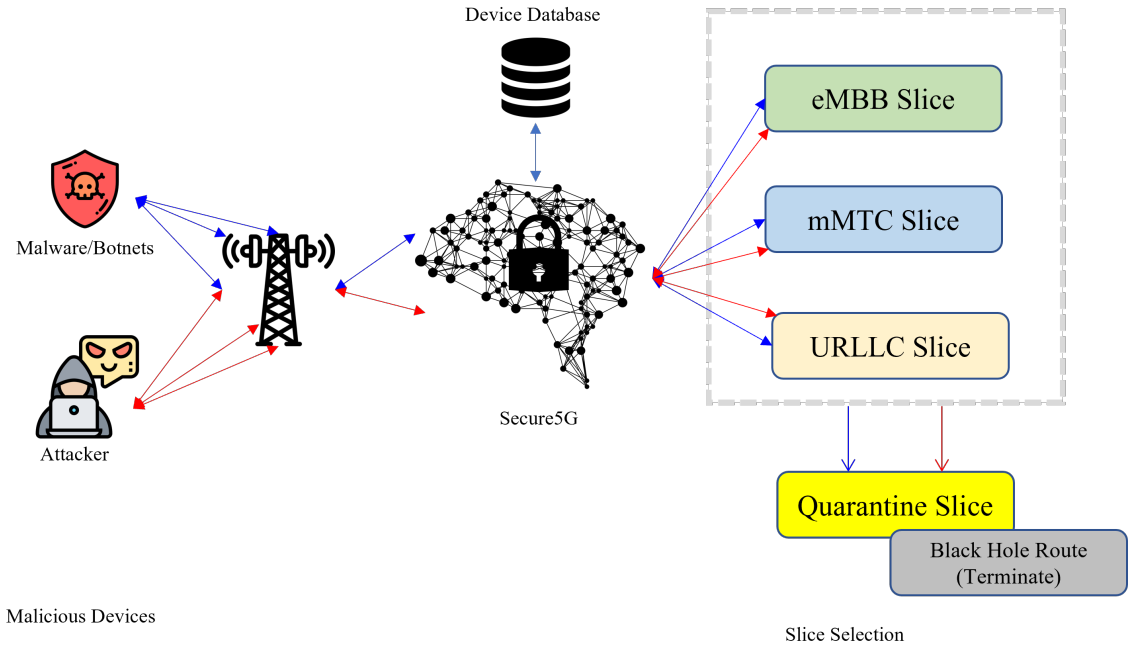

Figure 83: Volume-based Attack (Flooding) in Network Slicing

ing connections. Such devices, on the quarantine slice, will only get the bare minimum service at first, and eventually terminated if their malicious activity is confirmed.

The Secure5G model continues to learn and detect the incoming connections or the traffic pattern each time. If any known rogue (detected attacker) device continues its suspicious behavior by trying to flood the network, it will, ultimately, be denied any more service by the network. All its traffic will be moved to the black hole route and this device will be marked as a possible threat into our database. Any device, after an initial attach, can make an immediate connection request to a different slice. We would not want any genuine user or an authentic connection request to be flagged as a suspicious user just because they make multiple requests divided by a short time interval. Therefore, we consider a certain expiry timer on every incoming request to avoid the false flagging.

Secure5G model will only flag this to be a threat if the device makes multiple requests to different slices after the timer has expired. Timer expiry is customizable, and here we have considered 5 seconds in our simulation, but network operators can choose to define their threshold based on slice requirements.

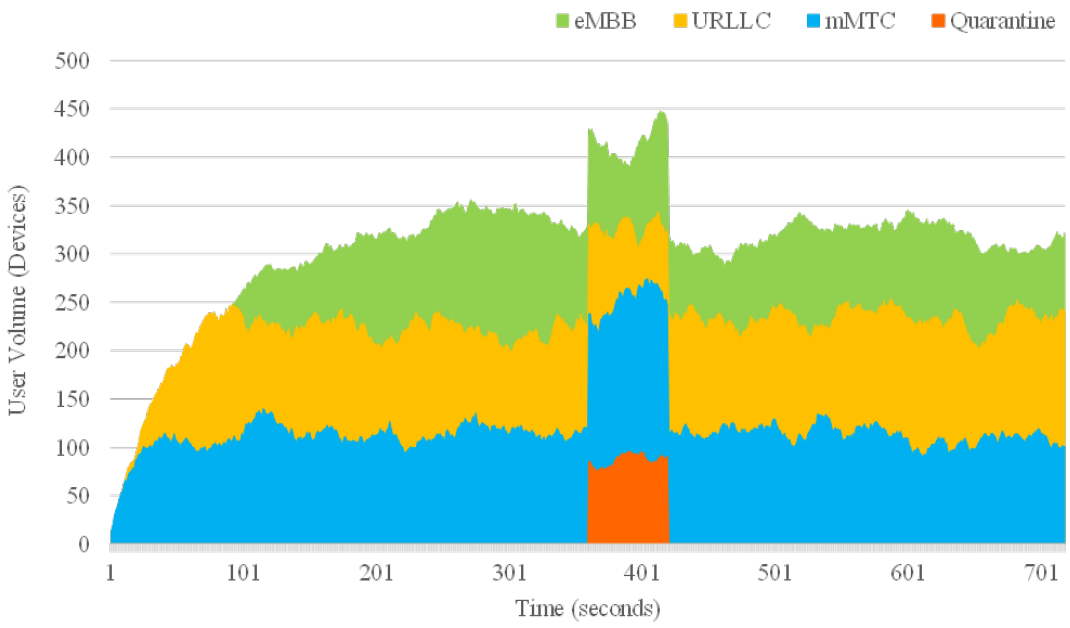

Figure 84: Volume-based Attack (Flooding)

Fig. 84 represents the volume-based cluster simulated using Secure5G providing the granular visibility of all incoming connection to different slices. The graph shows two scenarios combined; the normal and the attacker traffic. Normal traffic is when all devices follow the normal slice selection logic with DeepSlice, and the attack scenario shows the traffic with malicious connections where we used ten malicious devices, randomly, requesting resources from all slice at the same time. This is shown in the central region

of the plot as a sudden increase in the number of users. Our model will quickly identify this and transfer this malicious traffic to the quarantine slice

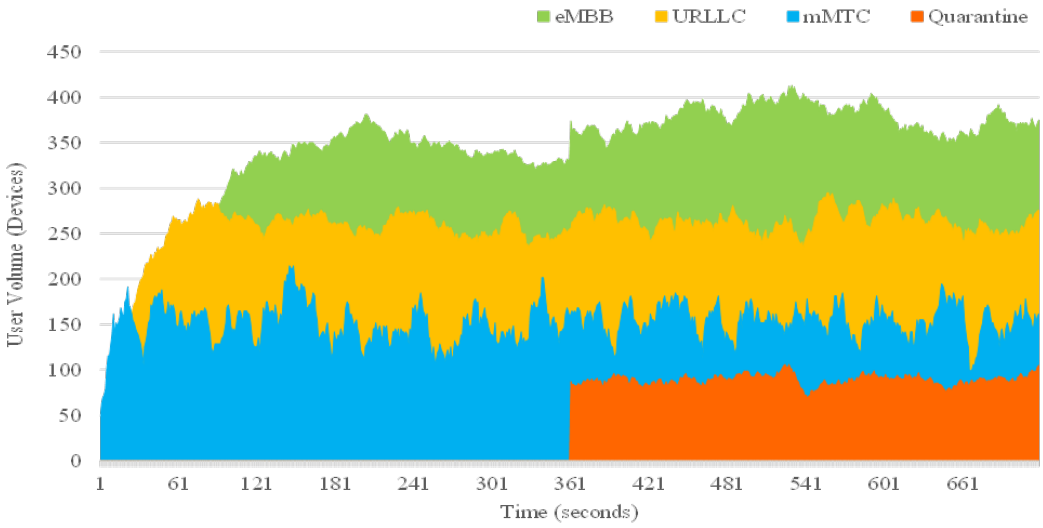

Figure 85: Slice-based Attack (Flooding by 'IoT')

Additionally, we evaluated the slice centric attack scenario in our simulation as shown in Fig. 85, where ten identified IoT devices are bombarding network with multiple request for resources from eMBB and URLLC slice instead of their standard mMTC. Secure5G will kick-in immediately after observing such a DDoS traffic stream as it identifies an influx of packets with the suspiciously-identical device, making multiple slice connections that do not match a typical pattern. By tracking such minuscule abnormalities, the Secure5G will weed out malicious traffic without impacting regular (genuine) user flow.

5.4.1.2 Masking Botnets (Spoofing) Attack

We have evaluated the robustness of our Secure5G model by simulating the spoofing attack scenario by hackers. In simple terms, device or client masking is an impersonation of a user where the attacker disguises the original source of an attack and allows the infected traffic to appear legitimate. Hackers commonly spoof DNS servers and IP addresses for spreading the virus. Botnets are malware infests devices which attacker uses to generate massive traffic to consume the server capacity, multiple network connection requests to flood the system resulting in server downtime, and in most scenarios, even without the knowledge of their owners.

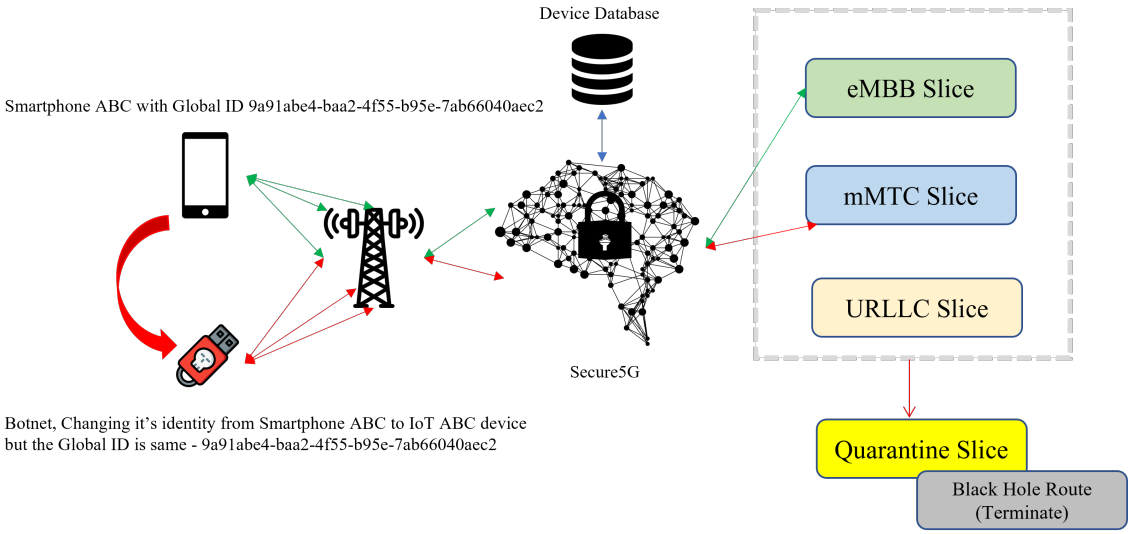

Figure 86: Device Masking Attack in Network Slicing

In Secure5G, we have simulated spoofing attacks by masking; for example, a smartphone device appears to be an IoT device and makes a slice request for mMTC instead of eMBB. Secure5G model has an inbuilt database of devices and user pattern

from learning, which maintains all original (and previous) connection requests made by any device. Secure5G assigns a unique global identifier when a UE tries to connect to the network first time. For example, when a smartphone device with certain International Mobile Equipment Identity (IMEI) e.g. 123456789012345 requests a connection, the Secure5G model will assign a GUTI (e.g. 9a91abe4-baa2-4f55-b95e-7ab66040aec2) to this IMEI and this information is stored in our database. Every GUTI is mapped with the device type, IMSI, slice requested, etc. If a hacker tries to mask this smartphone as an IoT device and requests mMTC slice, the model will understand this malicious request and flag it as possible botnet by comparing against IMEI and GUTI value in the database.

5.4.2 Prposed Work

We are expanding the scope of our 'Secure5G' model and below proposed ideas are work in progress:

5.4.2.1 Secure UE Capability

UE capability, RRC messages, and measurement reports [31] can be very helpful in identifying the false base stations. With Secure5G, proposed work is to utilize the UE Capability Enquiry and UE Capability Information messages along with the measurement of neighboring cells to identify the non-3GPP access or unauthorized base stations. Upon detection, the network can flag the false base station as a threat and direct another UE's to not connect to the identified false base station.

5.4.2.2 SIM Security in 5G

Article [137] presents a recent security incident on SIM card fraud where subscriber identity and property were stolen by hackers with a simple concept of SIM swapping. The network operator has failed to secure the subscriber account adequately. With Secure5G, a frequent change in device hardware, IMSI provisioning modification and unusual request to network access from different locations can be implemented, providing a much-secured ecosystem for device and SIM.

5.4.2.3 Unauthorized usage of shared resources between slice

The 3GPP specifications allow network operators to modify the RAN or Core network slice configurations for already deployed slice instances. Network operators can update (add/delete/modify) the network functions and change the security policies while it is in use. These flexibilities open a floodgate for potential threats to slice operations, and malicious actors can modify the QoS/SLA for a targeted slice. We are exploring areas to implement Secure5G in Core and RAN slicing to mitigate some of the in-network security threats.

5.4.2.4 Network Slice Exhaust

One attacker could potentially access the slice that could have lower-level security. For example, a slice for consumer will have lower security compared to the security measures for Industrial or Enterprise IoT, and malicious attacker can exhaust the resources in consumer slice. Though the slices are virtually isolated but if the network function resources are common to multiple slices (e.g. hardware resources: memory, processing

power or authentication). An ideal solution would be to pre-allocate security protocol resources for individual slices or ring-fenced resources in such a way that a slice has the individual capability to run irrespective of exhaustion on other slices [27].

5.5 Conclusion and Future Scope

Network slicing in 5G is a critical feature for next generation wireless networks, mobile operators and businesses. We have demonstrated the benefits of using DeepSlice for accurately predicting the best network slice based on device key parameters and orchestrated the handling of network load balancing and network slice failure using neural network models. Our future work will include emulating the developed model in a real production environment once the 5G ecosystem with devices and networks are commercially available for consumers. We will also extend and further improve this model to handle scenarios such as handovers, caching and predicting the future load, borrowing resources from other slices, and application-based slice management use cases.

This work also investigated the security concerns in the 5G network and presented a deep learning neural network model to create a robust Network Slicing framework to combat DDoS attacks filtering the malicious UE connections to the 5G network. Volume-based flooding and spoofing attack scenarios were used as illustrations to evaluate the overall performance, and the detection accuracy was more than 98% with our limited dataset. We believe the Secure5G implementation with DeepSlice will ensure the end-to-end security of the 5G network. We are considering several directions to improve further and extend the model to implement Secure5G into RAN, MEC, and Core Slicing. The

future model will also include the on-device and traffic behavior learning to train the model in real-time using reinforcement and recurrent learning, this will help us achieve more detection accuracy for secured 5G ecosystem.

CHAPTER 6

CONCLUSION AND FUTURE SCOPE

Cellular handovers are generally considered as simple or exponential arrivals; however, this fails to reflect the network and environmental consequences. Representing a handover connection arrival as a Matrix Exponential (ME) can help build efficient handover algorithms. Our approach regards handovers as ME. Single model can study all network characteristics and movement dynamics. Our model is simple and adaptable but captures all the handoff criteria. Future work will build formulations without renewal assumptions and compare B matrix formulations to real-world traffic, wireless signal propagation, and user movement dynamics. Next-generation wireless networks will contain several cell sites. Mobility management and user experience will be crucial for a service provider, and Deep Learning will be vital in building future autonomous and self-sufficient networks. We created a unique base model to learn and evaluate UE and network characteristics jointly to optimize user mobility. We use as many factors as needed to properly forecast handovers based on UE or network inputs. This will help develop 5G/6G handover algorithms that combine hyper-network densification, mmWave, M2M traffic, and ultra-reliable low latency communication.

ME models offer a full grasp of catastrophe recovery with repair models. Our work covered the entire formulation of the ME model, including matrix formulations, computations of the infinitesimal rate matrix, and transient probabilities incorporating Kronecker product dependencies. Our findings included transient and steady state analy-

ses of the ME model, a scalable survivability model, and service restoration and availability numbers. The ME model's survivability design examined the appropriate number of BSs and maintenance workers to enhance network availability. The survivability design section discussed how rapid and slow repairs affect total restoration time. It also showed how altering a single parameter affects the mending procedure. The chapter's insights will help any cellular service provider create a survivable network and restore it quickly with less resources. Future research might incorporate linked failures of macro cell BSs, small cell BSs, and backhaul links in the birth-death sequential Markov chain. Stochastic geometry might contribute geographic knowledge. When BSs fail, repairs might be prioritized in regions with poor coverage. In addition to failure analysis and temporal dynamics, financial analysis may help with survivable design. It might give cost models to optimize repair staff numbers and kinds vs network failure losses.

FDOP improves connection quality and demands earlier handovers than SINR. Usable coverage may be lower than projected, yet the handover procedure may take less time. With Fractional Packet Duplication, we can only duplicate what's needed. Since FDOP shows user and application quality, duplicated packets are reduced. It helps the network make an early decision and protects the user's network connection, meeting URLLC criteria. Our can be expanded to incorporate the Stochastic geometry of the surroundings to make FDOP more accurate. To satisfy 5G's dependability and low latency requirements, radio resources must be utilized effectively. Multi-Connectivity improves Spatial Diversity, beam shaping, and massive-MIMO for mmWave connections. Our Adaptive Fractional Packet Duplication technique allows the network to toggle PD ON and OFF.

Our multiple SINR or Fade threshold approaches are the most successful since they involve modest network algorithm adjustments. Since PD is inefficient during the whole transmission duration, our simulation findings illustrate when and where PD is effective and help understand channel circumstances. A network operator chooses PD based on resource availability and application needs. Future work can include more than two connections and can also include the WiFi6 standards to improve data rates and help with cellular network offloading.

5G network slicing is essential for next-generation wireless networks, mobile carriers, and businesses. DeepSlice correctly predicts the optimum network slice based on device key factors, and neural network model handles network load balancing and slice failure. Our work researched 5G network security and constructed a robust Network Slicing framework to mitigate DDoS assaults by screening malicious UE connections. Volume-based flooding and spoofing attack scenarios were employed to assess overall performance. With our limited dataset, detection accuracy was above 98%. Secure5G with DeepSlice will assure 5G network security end-to-end. We're investigating numerous ways to enhance and expand Secure5G into RAN, MEC, and Core Slicing. Future models will combine on-device and traffic behavior learning to train the model in real-time utilizing reinforcement and recurrent learning, improving detection accuracy for a secure 5G environment. Once 5G devices and networks are commercially accessible, we'll emulate the concept in a true production setting. We'll further expand this model to handle handovers, caching and forecasting future demand, borrowing resources from other slices, and application-based slice management.

REFERENCE LIST

[1] (2020) GSMA, the state of mobile internet connectivity. [Online]. Available: https://www.gsma.com/r/somic/

[2] J. G. Andrews, S. Buzzi, W. Choi, S. V. Hanly, A. Lozano, A. C. K. Soong, and J. C. Zhang, "What will 5G be?" *IEEE Journal on Selected Areas in Communications*, vol. 32, no. 6, pp. 1065–1082, 2014.

[3] "Ericsson 2018 report." [Online]. Available: https://www.ericsson.com/assets/local/mobility-report/documents/2018/ericsson-mobility-report-november-2018.pdf

[4] H. Baumann and W. Sandmann, "Numerical solution of level dependent quasi-birth-and-death processes," *Procedia Computer Science*, vol. 1, no. 1, pp. 1561–1569, 2010.

[5] T. Christensen, B. F. Nielsen, and V. B. Iversen, "Phase-type models of channel-holding times in cellular communication systems," *IEEE Transactions on Vehicular Technology*, vol. 53, no. 3, pp. 725–733, 2004.

[6] A. L. E. Corral-Ruiz, F. A. Cruz-Pérez, and G. Hernández-Valdez, "Coxian distribution modeling for the generalized and unified teletraffic analysis of mobile cellular networks," in *2010 7th International Conference on Electrical Engineering Computing Science and Automatic Control*, 2010, pp. 315–320.

[7] J. Zhou and C. Beard, "Comparison of combined preemption and queuing schemes for admission control in a cellular emergency network," in *IEEE Wireless Communications and Networking Conference, 2006. WCNC 2006.*, vol. 1, 2006, pp. 122–128.

[8] Zhou and C. Beard, "Balancing competing resource allocation demands in a public cellular network that supports emergency services," *IEEE Journal on Selected Areas in Communications*, vol. 28, no. 5, pp. 644–652, 2010.

[9] K. Mitchell, K. Sohraby, A. van de Liefvoort, and J. Place, "Approximation models of wireless cellular networks using moment matching," *IEEE Journal on Selected Areas in Communications*, vol. 19, no. 11, pp. 2177–2190, 2001.

[10] X. Wang and P. Fan, "Channel holding time in wireless cellular communications with general distributed session time and dwell time," *IEEE Communications Letters*, vol. 11, no. 2, pp. 158–160, 2007.

[11] J. Zhou and C. Beard, "A controlled preemption scheme for emergency applications in cellular networks," *IEEE Transactions on Vehicular Technology*, vol. 58, no. 7, pp. 3753–3764, 2009.

[12] G. Saraph and P. Singh, "New scheme for ip routing and traffic engineering," in *Workshop on High Performance Switching and Routing, 2003, HPSR.*, 2003, pp. 227–232.

[13] A. Zavabeti and M. Sarvi, "Reducing handover drops by utilizing comparative queuing," in *2009 Wireless Telecommunications Symposium*, 2009, pp. 1–5.

[14] F. Guidolin, I. Pappalardo, A. Zanella, and M. Zorzi, "A markov-based framework for handover optimization in hetnets," in *2014 13th Annual Mediterranean Ad Hoc Networking Workshop (MED-HOC-NET)*, 2014, pp. 134–139.

[15] J. Kumaran, K. Mitchell, and A. Van de Liefvoort, "Characterization of the departure process from an me/me/1 queue," *RAIRO-Operations Research*, vol. 38, no. 2, pp. 173–191, 2004.

[16] M. T. Gardner, C. C. Beard, and A. Van de Liefvoort, "Mission critical publish-subscribe performance modeling using linear algebraic and classical methods." in *SpringSim (TMS-DEVS)*, 2015, pp. 269–277.

[17] M. T. Gardner, C. Beard, and A. van de Liefvoort, "Efficient matrix-exponential random variate generation using a numeric linear combination approach," in *Proceedings of the Symposium on Theory of Modeling & Simulation-DEVS Integrative*, 2014, pp. 1–8.

[18] H.-W. Ferng and Y.-Y. Huang, "Handover scheme with enode-b pre-selection and parameter self-optimization for lte-a heterogeneous networks," in *2016 International Conference on Machine Learning and Cybernetics (ICMLC)*, vol. 2, 2016, pp. 594–599.

[19] V. Yajnanarayana, H. Rydén, and L. Hévizi, "5g handover using reinforcement learning," in *2020 IEEE 3rd 5G World Forum (5GWF)*, 2020, pp. 349–354.

[20] M. Erel-Özçevik and B. Canberk, "Road to 5g reduced-latency: A software defined handover model for embb services," *IEEE Transactions on Vehicular Technology*, vol. 68, no. 8, pp. 8133–8144, 2019.

[21] M. Alhabo, L. Zhang, N. Nawaz, and H. Al-Kashoash, "Game theoretic handover optimisation for dense small cells heterogeneous networks," *IET Communications*, vol. 13, no. 15, pp. 2395–2402, 2019.

[22] R. A. Paropkari, C. Beard, and A. Van De Liefvoort, "Handover performance prioritization for public safety and emergency networks," in *2017 IEEE 38th Sarnoff Symposium*, 2017, pp. 1–6.

[23] A. Alhammadi, M. Roslee, M. Y. Alias, I. Shayea, S. Alriah, and A. B. Abas, "Advanced handover self-optimization approach for 4g/5g hetnets using weighted fuzzy logic control," in *2019 15th International Conference on Telecommunications (ConTEL)*, 2019, pp. 1–6.

[24] A. Thantharate, R. Paropkari, V. Walunj, and C. Beard, "Deepslice: A deep learning approach towards an efficient and reliable network slicing in 5g networks," in *2019 IEEE 10th Annual Ubiquitous Computing, Electronics & Mobile Communication Conference (UEMCON)*, 2019, pp. 0762–0767.

[25] T. Bag, S. Garg, D. Preciado, Z. Shaik, J. Mueckenheim, and A. Mitschele-Thiel, "Self-Organizing Network functions for handover optimization in LTE Cellular networks," in *Mobile Communication - Technologies and Applications; 24. ITG-Symposium*. Berlin, Germany: VDE, May 2019, pp. 1–7. [Online]. Available: https://ieeexplore.ieee.org/document/8731768

[26] A. Thantharate, R. Paropkari, V. Walunj, C. Beard, and P. Kankariya, "Secure5g: A deep learning framework towards a secure network slicing in 5g and beyond," in *2020 10th Annual Computing and Communication Workshop and Conference (CCWC)*, 2020, pp. 0852–0857.

[27] R. A. Paropkari, A. A. Gebremichail, and C. Beard, "Fractional packet duplication and fade duration outage probability analysis for handover enhancement in 5g cellular networks," in *2019 International Conference on Computing, Networking and Communications (ICNC)*, 2019, pp. 298–302.

[28] S. L. Harja and Hendrawan, "Evaluation and optimization handover parameter based x2 in lte network," in *2017 3rd International Conference on Wireless and Telematics (ICWT)*, 2017, pp. 175–180.

[29] A. S. Priyadharshini and P. T. V. Bhuvaneswari, "A study on handover parameter optimization in lte-a networks," in *2016 International Conference on Microelectronics, Computing and Communications (MicroCom)*, 2016, pp. 1–5.

[30] L. M. Abdullah, M. D. Baba, and S. G. A. Ali, "Parameters optimization for handover between femtocell and macrocell in lte-based network," in *2014 IEEE In-*

ternational Conference on Control System, Computing and Engineering (ICCSCE 2014), 2014, pp. 636–640.

[31] A. Thantharate, C. Beard, and S. Marupaduga, "An approach to optimize device power performance towards energy efficient next generation 5g networks," in *2019 IEEE 10th Annual Ubiquitous Computing, Electronics & Mobile Communication Conference (UEMCON)*, 2019, pp. 0749–0754.

[32] V. Capdevielle, A. Feki, and A. Fakhreddine, "Self-optimization of handover parameters in LTE networks," in *2013 11th International Symposium and Workshops on Modeling and Optimization in Mobile, Ad Hoc and Wireless Networks (WiOpt)*. IEEE, May 2013, pp. 133–139. [Online]. Available: https://ieeexplore.ieee.org/document/6576426

[33] M. A. Ahmed, N. M. Ibrahim, and H. El-Tamally, "Performance evaluation of handoff queuing schemes," in *2010 Second International Conference on Communication Software and Networks*, 2010, pp. 83–87.

[34] L. Lipsky, *Queueing Theory: A Linear Algebraic Approach*. Springer New York, 2008. [Online]. Available: https://books.google.com/books?id=hyG8A_KMzesC

[35] "3gpp ts 36.300, evolved universal terrestrial radio access (e-utra) and evolved universal terrestrial radio access network (e-utran); overall description ; stage 2 (release 9), 9.5.0 ed., 2010." [Online]. Available: https://www.etsi.org/deliver/etsi_ts/136300_136399/136300/09.04.00_60/ts_136300v090400p.pdf

[36] W. Ni, I. B. Collings, and R. P. Liu, "Relay handover and link adaptation design for fixed relays in imt-advanced using a new markov chain model," *IEEE Transactions on Vehicular Technology*, vol. 61, no. 4, pp. 1839–1853, 2012.

[37] R. N. Clarke, "Expanding mobile wireless capacity: The challenges presented by technology and economics," *Telecommunications Policy*, vol. 38, no. 8, pp. 693–708, 2014. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0308596113001900

[38] (2019, April) 3GPP release-15. [Online]. Available: https://www.3gpp.org/release-15

[39] FirstNet, November 2020. [Online]. Available: https://firstnet.gov/system/tdf/FirstNet_Roadmap.pdf?file=1&type=node&id=1055&force=0

[40] A. Kwasinski, W. W. Weaver, P. L. Chapman, and P. T. Krein, "Telecommunications power plant damage assessment for hurricane katrina– site survey and follow-up results," *IEEE Systems Journal*, vol. 3, no. 3, pp. 277–287, 2009.

[41] T. Adachi, Y. Ishiyama, Y. Asakura, and K. Nakamura, "The restoration of telecom power damages by the great east japan earthquake," in *2011 IEEE 33rd International Telecommunications Energy Conference (INTELEC)*, 2011, pp. 1–5.

[42] K. K. Ghani Abbas, *Optical Networking Standards*. Springer US, 2006, ch. Network Survivability, pp. 295–319.

[43] Y. Ran, "Considerations and suggestions on improvement of communication network disaster countermeasures after the wenchuan earthquake," *IEEE Communications Magazine*, vol. 49, no. 1, pp. 44–47, 2011.

[44] K. T. Morrison, "Rapidly recovering from the catastrophic loss of a major telecommunications office," *IEEE Communications Magazine*, vol. 49, no. 1, pp. 28–35, 2011.

[45] Y. Liu, V. B. Mendiratta, and K. S. Trivedi, "Survivability analysis of telephone access network," in *15th International Symposium on Software Reliability Engineering*, 2004, pp. 367–377.

[46] A. Zolfaghari and F. J. Kaudel, "Framework for network survivability performance," *IEEE Journal on Selected Areas in Communications*, vol. 12, no. 1, pp. 46–51, 1994.

[47] L. Xie, P. E. Heegaard, and Y. Jiang, "Survivability analysis of a two-tier infrastructure-based wireless network," *Computer Networks*, vol. 128, pp. 28–40, 2017. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1389128617301226

[48] S. V. Dhople and A. D. Dominguez-Garcia, "A parametric uncertainty analysis method for markov reliability and reward models," *IEEE Transactions on Reliability*, vol. 61, no. 3, pp. 634–648, 2012.

[49] S. Dharmaraja, V. Jindal, and U. Varshney, "Reliability and survivability analysis for umts networks: An analytical approach," *IEEE Transactions on Network and Service Management*, vol. 5, no. 3, pp. 132–142, 2008.

[50] S. S. Gokhale and K. S. Trivedi, "Analytical models for architecture-based software reliability prediction: A unification framework," *IEEE Transactions on Reliability*, vol. 55, no. 4, pp. 578–590, 2006.

[51] L. Cui, Y. Xu, and X. Zhao, "Developments and applications of the finite markov chain imbedding approach in reliability," *IEEE Transactions on Reliability*, vol. 59, no. 4, pp. 685–690, 2010.

[52] A. Ghasemi, S. Yacout, and M. . Ouali, "Evaluating the reliability function and the mean residual life for equipment with unobservable states," *IEEE Transactions on Reliability*, vol. 59, no. 1, pp. 45–54, 2010.

[53] W.-J. Hsin and A. van de Liefvoort, "Analytical observations for a multiservice node," *Performance Evaluation*, vol. 51, no. 2, pp. 103–116, 2003.

[54] J. Kumaran, K. Mitchell, and A. van de Liefvoort, "Characterization of the departure process from an me/me/1 queue," *RAIRO - Operations Research*, vol. 38, no. 2, pp. 173–191, 2004.

[55] S. Neumayer, G. Zussman, R. Cohen, and E. Modiano, "Assessing the vulnerability of the fiber infrastructure to disasters," *IEEE/ACM Transactions on Networking*, vol. 19, no. 6, pp. 1610–1623, 2011.

[56] P. K. Agarwal, A. Efrat, S. K. Ganjugunte, D. Hay, S. Sankararaman, and G. Zussman, "Network vulnerability to single, multiple, and probabilistic physical attacks," in *2010 - MILCOM, 2010 Military Communications Conference*, 2010, pp. 1824–1829.

[57] M. Haenggi, *Stochastic Geometry for Wireless Networks*. Cambridge University Press, 2012.

[58] J. G. Andrews, F. Baccelli, and R. K. Ganti, "A tractable approach to coverage and rate in cellular networks," *IEEE Transactions on Communications*, vol. 59, no. 11, pp. 3122–3134, 2011.

[59] H. S. Dhillon, R. K. Ganti, F. Baccelli, and J. G. Andrews, "Modeling and analysis of k-tier downlink heterogeneous cellular networks," *IEEE Journal on Selected Areas in Communications*, vol. 30, no. 3, pp. 550–560, 2012.

[60] V. Jindal, S. Dharmaraja, and K. S. Trivedi, "Analytical survivability model for fault tolerant cellular networks supporting multiple services," *Simulation Series*, vol. 38, no. 3, p. 505, 2006.

[61] "Enhanced network survivability performance," 2021. [Online]. Available: https://ansi.org/

[62] P. E. Heegaard and K. S. Trivedi, "Network survivability modeling," *Computer Networks*, vol. 53, no. 8, pp. 1215–1234, Jun. 2009.

[63] M. L. Gamiz, "Smoothed estimation of a 3-state semi-markov reliability model," *IEEE Transactions on Reliability*, vol. 61, no. 2, pp. 336–343, 2012.

[64] D. Abusch-Magder, P. Bosch, T. E. Klein, P. A. Polakos, L. G. Samuel, and H. Viswanathan, "911-now: A network on wheels for emergency response and disaster recovery operations," *Bell Labs Technical Journal*, vol. 11, no. 4, pp. 113–133, 2007.

[65] M. Y. Selim and A. E. Kamal, "Post-disaster 4G/5G network rehabilitation using drones: Solving battery and backhaul issues," in *2018 IEEE Globecom Workshops (GC Wkshps)*, 2018, pp. 1–6.

[66] L. Zhong, K. Garlichs, S. Yamada, K. Takano, and Y. Ji, "Mission planning for uav-based opportunistic disaster recovery networks," in *2018 15th IEEE Annual Consumer Communications Networking Conference (CCNC)*, 2018, pp. 1–6.

[67] A. Merwaday and I. Guvenc, "Uav assisted heterogeneous networks for public safety communications," in *2015 IEEE Wireless Communications and Networking Conference Workshops (WCNCW)*, 2015, pp. 329–334.

[68] O. Hae-young. (2015) Korea unveils worlds first drone-based LTE service. [Online]. Available: http://www.koreaittimes.com/news/articleView.html

[69] A. Al-Hourani, S. Kandeepan, and A. Jamalipour, "Stochastic geometry study on device-to-device communication as a disaster relief solution," *IEEE Transactions on Vehicular Technology*, vol. 65, no. 5, pp. 3005–3017, 2016.

[70] A. Al-Hourani, S. Kandeepan, and E. Hossain, "Relay-assisted device-to-device communication: A stochastic analysis of energy saving," *IEEE Transactions on Mobile Computing*, vol. 15, no. 12, pp. 3129–3141, 2016.

[71] K. Trivedi, *Probability and statistics with reliability, queueing, and computer science applications (2nd ed.)*. Wiley-Interscience, 2001.

[72] R. A. Paropkari, C. Beard, and A. Van De Liefvoort, "Handover performance prioritization for public safety and emergency networks," in *2017 IEEE 38th Sarnoff Symposium*, 2017, pp. 1–6.

[73] M. Modarres, M. Kaminskiy, and V. Krivtsov, *Reliability Engineering and Risk Analysis: A Practical Guide (3rd ed.)*. CRC Press, 2016.

[74] J. Lorincz, L. Chiaraviglio, and F. Cuomo, "A measurement study of short-time cell outages in mobile cellular networks," *Computer Communications*, vol. 79, pp. 92–102, 2016.

[75] M. Al-Kuwaiti, N. Kyriakopoulos, and S. Hussein, "A comparative analysis of network dependability, fault-tolerance, reliability, security, and survivability," *IEEE Communications Surveys Tutorials*, vol. 11, no. 2, pp. 106–124, 2009.

[76] J. Murphy and T. W. Mogan, "Availability, reliability, and survivability: An introduction and some contractual implications," *The Journal of Defense Engineering*, pp. 26–29, 2006.

[77] J. Koroma, W. Li, and D. Kazakos, "A generalized model for network survivability," in *Proceedings of the 2003 conference on Diversity in computing*, 2003, pp. 47–51.

[78] L. Xie, P. E. Heegaard, and Y. Jiang, "Network survivability under disaster propagation: Modeling and analysis," in *2013 IEEE Wireless Communications and Networking Conference (WCNC)*, 2013, pp. 4730–4735.

[79] D. Öhmann, M. Simsek, and G. P. Fettweis, "Achieving high availability in wireless networks by an optimal number of rayleigh-fading links," in *2014 IEEE Globecom Workshops (GC Wkshps)*, 2014, pp. 1402–1407.

[80] T. T. Vu, L. Decreusefond, and P. Martins, "An analytical model for evaluating outage and handover probability of cellular wireless networks," in *The 15th International Symposium on Wireless Personal Multimedia Communications*. IEEE, Sep. 2012, pp. 643–647. [Online]. Available: https://ieeexplore.ieee.org/document/6398704

[81] A. M. Vegni, G. Tamea, T. Inzerilli, and R. Cusani, "A combined vertical handover decision metric for qos enhancement in next generation networks," in *2009 IEEE International Conference on Wireless and Mobile Computing, Networking and Communications*, 2009, pp. 233–238.

[82] H.-H. Choi, "An optimal handover decision for throughput enhancement," *IEEE Communications Letters*, vol. 14, no. 9, pp. 851–853, 2010.

[83] C.-N. Lee, S.-H. Lai, S.-C. Shen, and S.-I. Chen, "Multi-decision handover mechanism over wireless relay networks," in *2013 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference*, 2013, pp. 1–7.

[84] D. Öhmann and G. P. Fettweis, "Minimum duration outage of wireless rayleigh-fading links using selection combining," in *2015 IEEE Wireless Communications and Networking Conference (WCNC)*, 2015, pp. 681–686.

[85] D. Öhmann, A. Awada, I. Viering, M. Simsek, and G. P. Fettweis, "Sinr model with best server association for high availability studies of wireless networks," *IEEE Wireless Communications Letters*, vol. 5, no. 1, pp. 60–63, 2016.

[86] D. Ohmann, A. Awada, I. Viering, M. Simsek, and G. P. Fettweis, "Impact of mobility on the reliability performance of 5g multi-connectivity architectures," in *2017 IEEE Wireless Communications and Networking Conference (WCNC)*, 2017, pp. 1–6.

[87] F. B. Tesema, A. Awada, I. Viering, M. Simsek, and G. P. Fettweis, "Mobility modeling and performance evaluation of multi-connectivity in 5g intra-frequency networks," in *2015 IEEE Globecom Workshops (GC Wkshps)*, 2015, pp. 1–6.

[88] T. D. Novlan, R. K. Ganti, A. Ghosh, and J. G. Andrews, "Analytical evaluation of fractional frequency reuse for heterogeneous cellular networks," *IEEE Transactions on Communications*, vol. 60, no. 7, pp. 2029–2039, 2012.

[89] I. Viering, H. Martikainen, A. Lobinger, and B. Wegmann, "Zero-zero mobility: Intra-frequency handovers with zero interruption and zero failures," *IEEE Network*, vol. 32, no. 2, pp. 48–54, 2018.

[90] J. Lai and N. Mandayam, "Minimum duration outages in rayleigh fading channels," *IEEE Transactions on Communications*, vol. 49, no. 10, pp. 1755–1761, 2001.

[91] S. Nadarajah and S. Kotz, "Comments on 'minimum duration outages in rayleigh fading channels'," *IEEE Transactions on Communications*, vol. 55, no. 6, pp. 1110–1110, 2007.

[92] A. A. Gebremichail and C. Beard, "Fade duration based sleep mode activation in dense femtocell cluster networks," in *2017 26th International Conference on Computer Communication and Networks (ICCCN)*, 2017, pp. 1–6.

[93] Gebremichail and C. Beard, "Multi-hop relay selection based on fade durations," in *2015 Wireless Telecommunications Symposium (WTS)*, 2015, pp. 1–6.

[94] M. Tekinay, C. Beard, and A. van de Liefvoort, "Partial packet duplication: Control of fade and non-fade duration outages using matrix exponential distributions," *IEEE Transactions on Vehicular Technology*, vol. 69, no. 5, pp. 5652–5656, 2020.

[95] M. Mezzavilla, M. Zhang, M. Polese, R. Ford, S. Dutta, S. Rangan, and M. Zorzi, "End-to-end simulation of 5g mmwave networks," *IEEE Communications Surveys Tutorials*, vol. 20, no. 3, pp. 2237–2263, 2018.

[96] A. Assefa Gebremichail, C. Beard, and R. A. Paropkari, "Multi-hop relay selection based on fade durations," *Electronics*, vol. 9, no. 1, p. 92, 2020.

[97] T. Zugno, M. Polese, and M. Zorzi, "Integration of carrier aggregation and dual connectivity for the ns-3 mmwave module," in *Proceedings of the 10th Workshop on ns-3*, 2018, pp. 45–52.

[98] H. Kaja, R. A. Paropkari, C. Beard, and A. Van De Liefvoort, "Survivability and disaster recovery modeling of cellular networks using matrix exponential distributions," *IEEE Transactions on Network and Service Management*, vol. 18, no. 3, pp. 2812–2824, 2021.

[99] M. F. Özkoç, A. Koutsaftis, R. Kumar, P. Liu, and S. S. Panwar, "The impact of multi-connectivity and handover constraints on millimeter wave and terahertz cellular networks," *IEEE Journal on Selected Areas in Communications*, vol. 39, no. 6, pp. 1833–1853, 2021.

[100] R. A. Paropkari, A. Thantharate, and C. Beard, "Deep-mobility: A deep learning approach for an efficient and reliable 5g handover," in *2022 International Conference on Wireless Communications Signal Processing and Networking (WiSPNET)*, 2022, pp. 244–250.

[101] M. Giordani, M. Mezzavilla, S. Rangan, and M. Zorzi, "Multi-connectivity in 5g mmwave cellular networks," in *2016 Mediterranean Ad Hoc Networking Workshop (Med-Hoc-Net)*. IEEE, 2016, pp. 1–7.

[102] J. Rao and S. Vrzic, "Packet duplication for urllc in 5g: Architectural enhancements and performance analysis," *IEEE Network*, vol. 32, no. 2, pp. 32–40, 2018.

[103] P. Tinkhede and P. Ingole, "Survey of handover decision for next generation," in *International Conference on Information Communication and Embedded Systems (ICICES2014)*, 2014, pp. 1–5.

[104] P. Mishra, S. Kar, V. Bollapragada, and K.-C. Wang, "Multi-connectivity using nr-dc for high throughput and ultra-reliable low latency communication in 5g networks," in *2021 IEEE 4th 5G World Forum (5GWF)*. IEEE, 2021, pp. 36–40.

[105] V. Petrov, D. Solomitckii, A. Samuylov, M. A. Lema, M. Gapeyenko, D. Moltchanov, S. Andreev, V. Naumov, K. Samouylov, M. Dohler *et al.*, "Dynamic multi-connectivity performance in ultra-dense urban mmwave deployments," *IEEE Journal on Selected Areas in Communications*, vol. 35, no. 9, pp. 2038–2055, 2017.

[106] H. Zhang, W. Huang, and Y. Liu, "Handover probability analysis of anchor-based multi-connectivity in 5g user-centric network," *IEEE Wireless Communications Letters*, vol. 8, no. 2, pp. 396–399, 2018.

[107] M. Polese, M. Giordani, M. Mezzavilla, S. Rangan, and M. Zorzi, "Improved handover through dual connectivity in 5g mmwave mobile networks," *IEEE Journal on Selected Areas in Communications*, vol. 35, no. 9, pp. 2069–2084, 2017.

[108] E. J. Khatib, D. A. Wassie, G. Berardinelli, I. Rodriguez, and P. Mogensen, "Multi-connectivity for ultra-reliable communication in industrial scenarios," in *2019*

IEEE 89th Vehicular Technology Conference (VTC2019-Spring). IEEE, 2019, pp. 1–6.

[109] B. Kharel, O. L. A. López, N. H. Mahmood, H. Alves, and M. Latva-Aho, "Fog-ran enabled multi-connectivity and multi-cell scheduling framework for ultra-reliable low latency communication," *IEEE Access*, vol. 10, pp. 7059–7072, 2022.

[110] A. Aijaz, "Packet duplication in dual connectivity enabled 5g wireless networks: Overview and challenges," *IEEE Communications Standards Magazine*, vol. 3, no. 3, pp. 20–28, 2019.

[111] M.-T. Suer, C. Thein, H. Tchouankem, and L. Wolf, "Multi-connectivity as an enabler for reliable low latency communications—an overview," *IEEE Communications Surveys & Tutorials*, vol. 22, no. 1, pp. 156–169, 2019.

[112] F. B. Tesema, A. Awada, I. Viering, M. Simsek, and G. P. Fettweis, "Mobility modeling and performance evaluation of multi-connectivity in 5g intra-frequency networks," in *2015 IEEE Globecom Workshops (GC Wkshps)*. IEEE, 2015, pp. 1–6.

[113] S. J. Su and B. S. Kwon, "Performance analysis of packet duplication for reliability enhancement of wireless link," in *2019 International Conference on Information and Communication Technology Convergence (ICTC)*. IEEE, 2019, pp. 825–829.

[114] M. Polese, M. Mezzavilla, and M. Zorzi, "Performance comparison of dual connectivity and hard handover for lte-5g tight integration," *arXiv preprint arXiv:1607.05425*, 2016.

[115] A. Wolf, P. Schulz, M. Dörpinghaus, J. C. S. Santos Filho, and G. Fettweis, "How reliable and capable is multi-connectivity?" *IEEE Transactions on Communications*, vol. 67, no. 2, pp. 1506–1520, 2018.

[116] N. H. Mahmood, A. Karimi, G. Berardinelli, K. I. Pedersen, and D. Laselva, "On the resource utilization of multi-connectivity transmission for urllc services in 5g new radio," in *2019 IEEE Wireless Communications and Networking Conference Workshop (WCNCW)*. IEEE, 2019, pp. 1–6.

[117] P. Popovski, Č. Stefanović, J. J. Nielsen, E. De Carvalho, M. Angjelichinoski, K. F. Trillingsgaard, and A.-S. Bana, "Wireless access in ultra-reliable low-latency communication (urllc)," *IEEE Transactions on Communications*, vol. 67, no. 8, pp. 5783–5801, 2019.

[118] "Department of transportation, federal aviation administration (faa), "national airspace system requirements document," august 11, 2014."

[119] "3gpp release-15." [Online]. Available: https://www.3gpp.org/release-15

[120] S. O. Oladejo and O. E. Falowo, "5g network slicing: A multi-tenancy scenario," in *2017 Global Wireless Summit (GWS)*, 2017, pp. 88–92.

[121] L. Ma, X. Wen, L. Wang, Z. Lu, and R. Knopp, "An sdn/nfv based framework for management and deployment of service based 5g core network," *China Communications*, vol. 15, no. 10, pp. 86–98, 2018.

[122] P. Du and A. Nakao, "Deep learning-based application specific ran slicing for mobile networks," in *2018 IEEE 7th International Conference on Cloud Networking (CloudNet)*, 2018, pp. 1–3.

[123] R. Abhishek, S. Zhao, and D. Medhi, "Spartacus: Service priority adaptiveness for emergency traffic in smart cities using software-defined networking," in *2016 IEEE International Smart Cities Conference (ISC2)*, 2016, pp. 1–4.

[124] T. Yoo, "Network slicing architecture for 5g network," in *2016 International Conference on Information and Communication Technology Convergence (ICTC)*, 2016, pp. 1010–1014.

[125] F. Kurtz, C. Bektas, N. Dorsch, and C. Wietfeld, "Network slicing for critical communications in shared 5g infrastructures - an empirical evaluation," in *2018 4th IEEE Conference on Network Softwarization and Workshops (NetSoft)*, 2018, pp. 393–399.

[126] R. Abhishek, D. Tipper, and D. Medhi, "Network virtualization and survivability of 5g networks: Framework, optimization model, and performance," in *2018 IEEE Globecom Workshops (GC Wkshps)*, 2018, pp. 1–6.

[127] V. K. Choyi, A. Abdel-Hamid, Y. Shah, S. Ferdi, and A. Brusilovsky, "Network slice selection, assignment and routing within 5g networks," in *2016 IEEE Conference on Standards for Communications and Networking (CSCN)*, 2016, pp. 1–7.

[128] C. Campolo, A. Molinaro, A. Iera, R. R. Fontes, and C. E. Rothenberg, "Towards 5g network slicing for the v2x ecosystem," in *2018 4th IEEE Conference on Network Softwarization and Workshops (NetSoft)*, 2018, pp. 400–405.

[129] R. A. Addad, M. Bagaa, T. Taleb, D. L. C. Dutra, and H. Flinck, "Optimization model for cross-domain network slices in 5g networks," *IEEE Transactions on Mobile Computing*, vol. 19, no. 5, pp. 1156–1169, 2020.

[130] D. Sattar and A. Matrawy, "Towards secure slicing: Using slice isolation to mitigate ddos attacks on 5g core network slices," in *2019 IEEE Conference on Communications and Network Security (CNS)*, 2019, pp. 82–90.

[131] P. Schneider, C. Mannweiler, and S. Kerboeuf, "Providing strong 5g mobile network slice isolation for highly sensitive third-party services," in *2018 IEEE Wireless Communications and Networking Conference (WCNC)*, 2018, pp. 1–6.

[132] J. Ni, X. Lin, and X. S. Shen, "Efficient and secure service-oriented authentication supporting network slicing for 5g-enabled iot," *IEEE Journal on Selected Areas in Communications*, vol. 36, no. 3, pp. 644–657, 2018.

[133] A. Thantharate, C. Beard, and P. Kankariya, "Coap and mqtt based models to deliver software and security updates to iot devices over the air," in *2019 Interna-*

tional Conference on Internet of Things (iThings) and IEEE Green Computing and Communications (GreenCom) and IEEE Cyber, Physical and Social Computing (CPSCom) and IEEE Smart Data (SmartData), 2019, pp. 1065–1070.

[134] S. Rezvy, Y. Luo, M. Petridis, A. Lasebae, and T. Zebin, "An efficient deep learning model for intrusion classification and prediction in 5g and iot networks," in *2019 53rd Annual Conference on Information Sciences and Systems (CISS)*, 2019, pp. 1–6.

[135] "Attack landscape h1 2019: Iot, smb traffic abound." [Online]. Available: https://blog.f-secure.com/attack-landscape-h1-2019-iot-smb-traffic-abound

[136] "Deepslice dataset, (last accessed date 09.25.2019) kpi's used for training and testing the model." [Online]. Available: https://github.com/adtmv7/DeepSlice

[137] "At&t faces new usd 1.8 million lawsuit over sim hijacking attack." [Online]. Available: https://www.vice.com/en_us/article/3kxy7w/atandt-faces-new-dollar18-million-lawsuit-over-sim-hijacking-attack

VITA

Rahul Arun Paropkari was born on October 23, 1985 in Aurangabad, Maharashtra, India. He received his under-graduate degree from Government Engineering College at Aurangabad in 2007. In the August of 2009, he came overseas to pursue his masters in Electrical Engineering followed by a second masters in Computer Science followed by the Interdisciplinary - PhD degree from University of Missouri-Kansas City.

He has been employed full time in the telecommunications industry since 2007 and working on his research for the past seven years. His research has taken several turns over the years, but mainly focused on 4G-5G mobility, handover optimization, network resiliency and survivability modelling using queueing theory, deep learning and other ML related modelling for network slicing, cellular handovers, etc. Prior to graduation this year, he intends to complete his current work on cell selection, multi connectivity, and database creation tool to help other students efficiently train their deep learning models.

He has over 13 years of experience starting with Siemens enterprise communications (2007-2009) until he joined Innovating Solutions/Tappecue (2011-2012) designing and building the first Wi-Fi barbecue thermometer. He then joined Alcatel-Lucent/Nokia (2012-2015) to enter the cellular industry. Since then, he's had the opportunity to be with Sprint (2015-2018) working with multiple OEMs like Nokia, Samsung, Ericsson. He moved to the Product Development team within Sprint (2018-2020) managing the labs and later moved to his current position at T-Mobile (2020-Present) focusing on the 5G new feature design, development and deployment watching over the end-to-end quality.