UNIVERSITY OF OKLAHOMA

GRADUATE COLLEGE

A PARADIGM SHIFTING APPROACH IN SON FOR FUTURE CELLULAR NETWORKS

A DISSERTATION

SUBMITTED TO THE GRADUATE FACULTY

in partial fulfillment of the requirements for the

Degree of

DOCTOR OF PHILOSOPHY

By

HASAN FAROOQ
Norman, Oklahoma
2018

A PARADIGM SHIFTING APPROACH IN SON FOR FUTURE CELLULAR NETWORKS

A DISSERTATION APPROVED FOR THE
SCHOOL OF ELECTRICAL AND COMPUTER ENGINEERING

BY

Dr. Ali Imran, Chair

Dr. Thordur Runolfsson

Dr. Hazem Refai

Dr. Kam Wai Clifford Chan

Dr. Curt Adams

Acknowledgments

First and foremost, I am greatly thankful to Allah the Almighty for bestowing His countless blessings on me and giving me strength and courage for successful completion of this research work. I would like to express my deepest gratitude and appreciation to my supervisor Dr. Ali Imran who gave me opportunity to carry out this research work. It has been an honor to be his first Ph.D. student. He has taught me, both consciously and unconsciously, how a good research is done. I appreciate all his contributions of time, ideas, and funding to make my Ph.D. experience productive and stimulating. He was always there whenever I needed him. He gave me complete freedom to try out new ideas and continuously supported me throughout this research. The following quote describing best leader always reminds me of him "***A leader is best when people barely know he exists, not so good when people obey and acclaim him, worse when they despise him. But of a good leader who talks little when his work is done, his aim fulfilled, they will say: We did it ourselves***." I particularly appreciate his great patience for deeply reviewing my research work. I am also grateful to my doctoral dissertation committee members as well as to Dr. Muhammad Ali Imran and Dr. Pramode Verma who gave me instructive suggestions for improving this research work throughout the Ph.D. pursuit. I also owe a great debt of gratitude to Dr. Low Tang Jung, my MSc research supervisor in UTP Malaysia, for introducing me to the world of research. I would also like to thank my fellow colleagues in AI4Networks Laboratory@OU and Dr. Ahmed Zoha from 5GIC, University of Surrey for their support and providing a conducive research and work environment. Here I would like to mention Renee Wagenblatt and Krista Pettersen who took care of all administrative tasks and were of immense help throughout my four years of study at OU-Tulsa.

Lastly, I would like to thank my family for all their love and encouragement. I owe

my deepest gratitude to my parents and four loving sisters who gave me endless support and love throughout my life. And most of all for my loving, supportive, encouraging, and patient wife whose faithful support during the final stages of this Ph.D. is so appreciated. Thank you.

Table of Contents

List of Figures

List of Tables

Abstract

The race to next generation cellular networks is on with a general consensus in academia and industry that massive densification orchestrated by self-organizing networks (SONs) is the cost-effective solution to the impending mobile capacity crunch. While the research on SON commenced a decade ago and is still ongoing, the current form (i.e., the reactive mode of operation, conflict-prone design, limited degree of freedom and lack of intelligence) hinders the current SON paradigm from meeting the requirements of 5G. The ambitious quality of experience (QoE) requirements and the emerging multifarious vision of 5G, along with the associated scale of complexity and cost, demand a significantly different, if not totally new, approach to SONs in order to make 5G technically as well as financially feasible. This dissertation addresses these limitations of state-of-the-art SONs. It first presents a generic low-complexity optimization framework to allow for the agile, on-line, multi-objective optimization of future mobile cellular networks (MCNs) through only top-level policy input that prioritizes otherwise conflicting key performance indicators (KPIs) such as capacity, QoE, and power consumption. The hybrid, semi-analytical approach can be used for a wide range of cellular optimization scenarios with low complexity. The dissertation then presents two novel, user-mobility, prediction-based, proactive self-optimization frameworks (AURORA and OPERA) to transform mobility from a challenge into an advantage. The proposed frameworks leverage mobility to overcome the inherent reactiveness of state-of-the-art self-optimization schemes to meet the extremely low latency and high QoE expected from future cellular networks vis-à-vis 5G and beyond. The proactiveness stems from the proposed frameworks' novel capability of utilizing past hand-over (HO) traces to determine future cell loads instead of observing changes in cell loads passively and then reacting to them. A semi-Markov renewal process is leveraged to build a model that can predict the cell of the next HO and the time of the

HO for the users. A low-complexity algorithm has been developed to transform the predicted mobility attributes to a user-coordinate level resolution. The learned knowledge base is used to predict the user distribution among cells. This prediction is then used to formulate a novel (i) proactive energy saving (ES) optimization problem (AURORA) that proactively schedules cell sleep cycles and (ii) proactive load balancing (LB) optimization problem (OPERA). The proposed frameworks also incorporate the effect of cell individual offset (CIO) for balancing the load among cells, and they thus exploit an additional ultra-dense network (UDN)-specific mechanism to ensure QoE while maximizing ES and/or LB. The frameworks also incorporates capacity and coverage constraints and a load-aware association strategy for ensuring the conflict-free operation of ES, LB, and coverage and capacity optimization (CCO) SON functions. Although the resulting optimization problems are combinatorial and NP-hard, proactive prediction of cell loads instead of reactive measurement allows ample time for combination of heuristics such as genetic programming and pattern search to find solutions with high ES and LB yields compared to the state of the art. To address the challenge of significantly higher cell outage rates in anticipated in 5G and beyond due to higher operational complexity and cell density than legacy networks, the dissertation's fourth key contribution is a stochastic analytical model to analyze the effects of the arrival of faults on the reliability behavior of a cellular network. Assuming exponential distributions for failures and recovery, a reliability model is developed using the continuous-time Markov chains (CTMC) process. Unlike previous studies on network reliability, the proposed model is not limited to structural aspects of base stations (BSs), and it takes into account diverse potential fault scenarios; it is also capable of predicting the expected time of the first occurrence of the fault and the long-term reliability behavior of the BS.

The contributions of this dissertation mark a paradigm shift from the reactive,

semi-manual, sub-optimal SON towards a conflict-free, agile, proactive SON. By paving the way for future MCN's commercial and technical viability, the new SON paradigm presented in this dissertation can act as a key enabler for next-generation MCNs.

List of Key Symbols and Acronyms

b, s, r, q, c, u	as subscript or superscript denote association to base station, sector, small cell, q^{th} bin, cell or user respectively
Q_b	set of bins in which BS are located, $\mathcal{Q}_b \subseteq Q$
S	set of all sectors in the systems,where $\|\mathcal{S}\| = S$
S_b	total number of sectors b^{th} BS has
h_s	(antenna) height of s^{th} sector antenna on BS
H_r	set of all SC antenna heights
Υ_f	number of times spectrum is reused within site
R_b	number of SC in b^{th} BS
R	$\mathcal{R} = \{R_1, R_2, R_3 ... R_B\}$, $R = \|\mathcal{R}\| = \sum_{b=1}^{B} R_b$
ϕ^s	azimuth angle of s^{th} sector
θ	set of tilt angles of all sectors
θ^s	tilt angle of s^{th} sector
P^s	set of transmission powers of all sectors
p^s	transmission power from s^{th} sector
G_q^s	gain from the s^{th} sector antenna to q^{th} bin
α	path loss co-efficient including long term shadowing
β	path loss exponent
δ_q^s	shadowing from s^{th} to q^{th} bin
φ_v	vertical beam width of the antenna
φ_h	horizontal beam width of the antenna
Υ	capacity wise KPI
Ω	power consumption wise KPI
Λ	service area fairness wise KPI
$\mathcal{X}\backslash y$	means all elements of $\mathcal{X}$ except y
ψ	semi-Markov kernel
$\Gamma_{i,j}^{(u)}(t)$	sojourn time distribution of user u in cell i when next cell is j
$\Delta^{(u)}$	steady-state distribution of semi-Markov model for user u
$\xi_j^{(u)}$	mean sojourn time of user u in cell j
$\mathbb{N}_i$	set of all neighboring cells of cell i
$\mathbb{C}_N^u$	next probable cell of user u
$\mathbb{T}_{HO}^u$	time of next HO for user u
l_k^u	UE's current location coordinates at time instant k
$\chi_{i,m}^{(u)}$	semi-Markov state transient
π^c	indicator variable that will be 1(0) for ON(OFF) BS in cell c
η_c	cell load
γ_u^c	SINR achievable by user u in cell c
ω_B^u	bandwidth assigned to user u
$\hat{\tau}_u$	minimum required rate of the user
$\mathbb{U}_c$	number of active users connected to a cell c
N_{us}	number of unsatisfied or dropped users

$P^c_{r,u}$	RSRP received by user u in cell c
$\bar{\omega}$	coverage probability
η_T	cell load threshold
P^c_{CIO}	CIO offset for cell c
N_{PRB}	total number of PRBs in a cell
κ	noise variable
a	user association priority variable
k'	prediction interval
λ_c	critical failures arrival rate
λ_t	trivial failures arrival rate
μ_{dc}	average time to move the network from sub-optimal state back to optimal state
μ_c	average time to move the network from outage state back to optimal state
$\mathbf{G}$	generator matrix
$\mathbf{V}$	rate matrix
$\mathbf{I}$	identity matrix
ρ	uniform rate parameter
$\Psi_{i,j}(T)$	expected amount of time the CTMC spends in state j during the interval $[0, T]$, starting in state i
$\zeta_{i \to j}$	first passage time to state j starting from state i
MCN	mobile cellular network
LTE	long term evolution
$3GPP$	third generation partnership
UE	user equipment
SON	self-organizing network
PM	performance metric
COP	configuration and optimization parameters
MDT	minimization of drive tests
QoS	quality of service
KPI	key performance indicator
EE	energy efficiency
ES	energy saving
LB	load balancing
CCO	coverage and capacity optimization
$HetNet$	heterogeneous network
CIO	cell individual offset
UDN	ultra-dense network
QoE	quality of experience
SC	small cell
HO	hand over
$CTMC$	continuous time markov chain
BS	base station
$RSRP$	reference signal received power
$RSRQ$	reference signal received quality

RRC	radio resource control
CDR	call data records
CS	cellular system
CSO	cellular system optimization
NTC	network topology configurations
MCE	modulation coding efficiency
MCS	modulation coding schemes
$EIRP$	effective isotropic radiated power
ECR	energy consumption ratio
ERG	energy reduction gain
PRB	physical resource block
QCI	QoS class identifier
GA	genetic algorithm
PS	pattern search algorithm
OPC	optimization parameters configuration

CHAPTER 1

Introduction

1.1 Motivation for Paradigm Shift in Self-Organizing Networks

This century has witnessed an exponential increase in mobile data usage—around 400-million-fold over the past 15 years—thanks to the proliferation of smart devices and diversity in mobile applications. According to the latest visual network index report from Cisco [1], global mobile traffic will rise from 7.2 Exabytes per month in 2016 to reach 49.0 Exabytes per month by 2021. This has prompted the need for the future generation of wireless networks to provide unprecedented capacity gain and a top notch quality of service (QoS). The ambitious requirements of zero latency and gigabit experience are driving the evolution of fifth-generation (5G) cellular networks. While the surge in mobile device-based applications is only bounded by imagination, the capacity of cellular systems is tightly bounded by fundamental physics. The general consensus is that major capacity gain in 5G must come primarily from impromptu network densification. It is not difficult to prognosticate that such a colossal deployment will become a significant challenge in 5G aggravating several problems in terms of energy consumption, mobility management, and OPEX, to name a few. This means that automation of the post-deployment operation and optimization in MCN for reducing costs, handling complexity, and maximizing resource efficiency will not only become a necessity, but the future MCN's technical and commercial viability may also hinge on them.

The research on the automation of MCN operation and optimization commenced a decade ago in the context of self-organizing networks (SONs), and it is still ongoing. However, legacy SON solutions aim to only automate the manual process of opti-

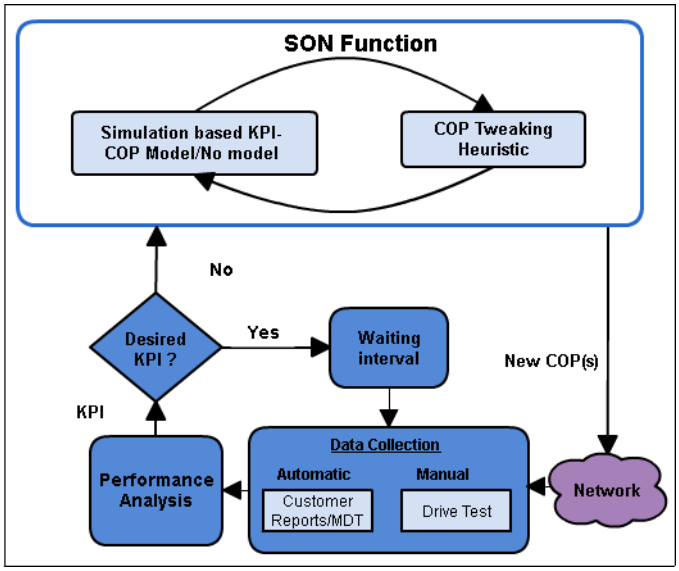


Fig. 1.1: Conventional SON Architecture

mizing one of the many performance metrics (PMs) using one or a few configuration and optimization parameters (COPs). Typical COP examples include antenna tilts, azimuths, Tx powers, frequency reuse, handover hysteresis, cell individual offsets (CIOs), cell discovery rate, and neighbor lists. Each PM-COP self-optimization routine is called a SON function. Driven by the dire need to cut the cost of manual operation, many SON functions have recently been standardized by 3GPP. The operation of a typical SON function is illustrated in Fig. 1.1 and can be explained as follows: performance data, which contains PMs mostly related to coverage and QoS are gathered through drive tests, customer complaints or minimization of drive test (MDT) reports. These data are analyzed to determine whether the PM under consideration is falling below a preset threshold. If so, then a SON function that handles that PM (e.g., coverage and capacity optimization [CCO] if the PM is either throughput or coverage) is triggered. To restore the PM, the SON function performs a hit-and-trial-based gradual adjustment of the COPs in the live network. The PM is observed after each COP adjustment and the process is repeated until a satisfactory PM is achieved. This legacy SON falls short of the mark for 5G requirements due to the following limitations:

1. **Legacy SON lacks intelligence:** The legacy SON relies on hit and trial to adjust the COPs, since no explicit model of the optimization objective as a function of the optimization parameters is derived, and the COP adjustment is performed through a basic hit-and-trial heuristic, with a goal of achieving improvement over the existing configuration without developing robust model to project the effect of a COP change on the system performance. While this hit-and-trial scheme can improve the performance, it may never optimize or maximize it.

2. **Legacy SON lacks ability to function with top level instructions:** Legacy SON functions are effectively stand-alone control loops that simply aim to eliminate mundane tasks that were previously performed by humans. The day-to-day operation of current MCNs involves tweaking a myriad of COPs such as antenna tilts, azimuths, Tx powers, frequency reuse, handover hysteresis, CIOs, cell discovery frequencies, and neighbor lists. Such a COP adjustment is done by engineers while relying on their domain knowledge. In some cases, this domain knowledge is aided by an offline system-level simulator. The goal of this laborious and almost continuous COP adjustment process is to enhance or maintain performance indicators that may ultimately improve one or more of the three main top-level key performance indicators (KPIs) of MCN, namely, capacity, QoS, and energy efficiency (EE). This manual process is known to be highly inefficient and prone to human error, and it is bound to become infeasible altogether in future MCN.

3. **Reactive mode of operation:** The plethora of existing SON approaches are designed to kick in after detecting network conditions that have already taken effect. For example, when load imbalance is detected in a network, a non-convex NP-hard load balancing (LB) algorithm is usually solved to optimize hard or soft network parameters. This is an improvement on fixed parameter settings in real networks that achieve LB at the cost of QoS. However, given the acute dynamics in HetNets, by the time a load imbalance is detected and a realistic non-convex NP-hard LB

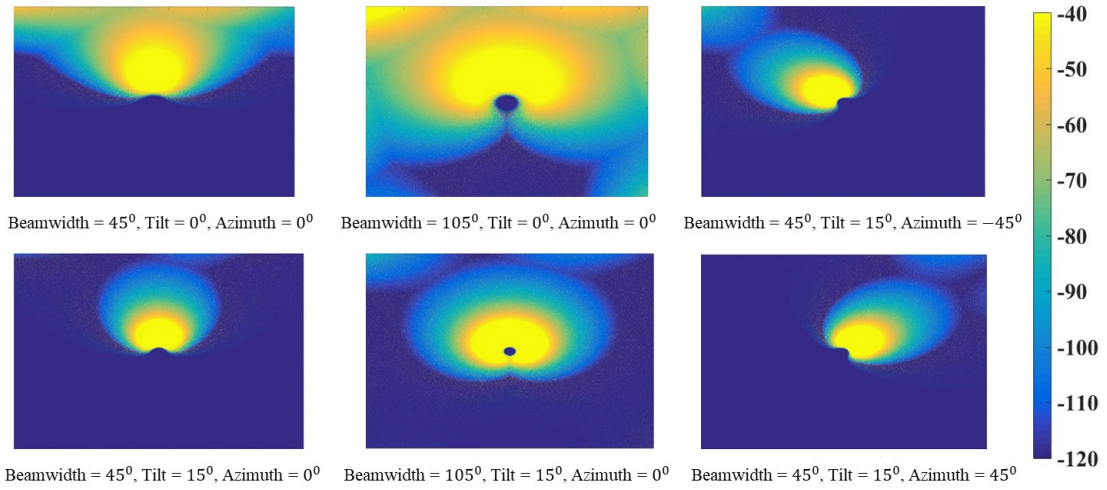


Fig. 1.2: Reference signal received power heatmap corresponding to various combination of beam widths, tilts and azimuth orientation of BS antennas with BS located at center of the square region

algorithm is solved to produce a new network configuration that is optimal for the observed network conditions, the conditions may have already changed. The newly determined optimal parameter settings are thus likely to be suboptimal before they can be actuated. This problem can be exacerbated, particularly in 5G, where myriad services and a plethora of cell types mean that the dynamics of the cellular eco-system will be even more swift.

4. **Limited set of optimization parameters:** Downlink transmission power is one of the prime optimization parameters that has been largely used in literature as an actuator for SON functions. However, with the evolution of smart antenna technology, a new set of optimization parameters has surfaced that is yet to be exploited. This includes beam widths (radiation patterns) that can be adapted on the fly by optimizing the phases of complex weight vectors—thanks to smart antenna technology. Similarly, the azimuth orientation of the antennas can be leveraged to effectively change the cell footprint in conjunction with the antenna tilts, as illustrated in Fig. 1.2. As per the Sobol-based variance sensitivity analysis method [2], the first-order sensitivity index values for some of the optimization

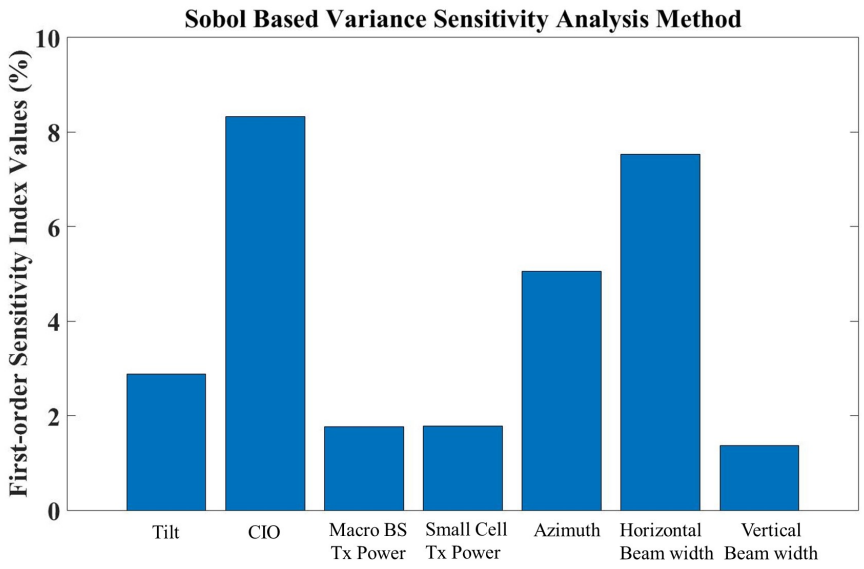


Fig. 1.3: Sobol method-based first-order sensitivity index values for tilts, CIOs, macro BS transmission power, small BS transmission power, azimuth, horizontal and vertical beam widths

parameters are plotted in Fig. 1.3. It is observed that the CIOs, horizontal beam width and azimuth are found to have the largest impact on network performance (the QoS). This observation calls for a deviation from the legacy age old paradigm of only optimizing Tx power to maximize system performance and keeping other control knobs untouched.

5. **Conflict-prone design of SONs:** One caveat with conventional SON solutions is that they are oblivious to the fact that multiple SON functions may be prone to hidden or undesired conflict when implemented together in a network [3]; e.g., the CCO SON use case may try to improve coverage by increasing Tx power, which in turn can force a large number of users to jump into its coverage thereby conflicting with LB SON objective. The interplay between CCO and LB becomes complicated, considering that they both resort to the optimization of the same parameters. The CIO, which unlike antenna parameters, is a soft parameter, was later introduced for LB and traffic steering in heterogeneous networks (HetNets). However, adjustment of the CIO by the LB algorithm may also cause conflict with CCO objectives as a

user offloaded due to increased CIO may face higher interference (assuming intra-frequency offloading), and lower received power from the destination cell, compared to the origin cell. This may result in lower SINR and ultimately lower throughputs. As explicated in [3], such a conflict-prone LB solution design can actually degrade a network's performance instead of improving it.

6. **Overly simplified, unrealistic assumptions:** Most of the existing literature on SONs is more theoretical in nature, using analytical models that often involve overly-simplified, unrealistic assumptions – such as uniformly distributed user equipments (UEs); a spatially independent distribution of base stations (BSs); omnidirectional, single-antenna transmission and reception; fixed transmit powers; the same CIO for all cells in one tier; and full load scenarios – to achieve convexity in the optimization problem. These assumptions help to make the analysis tractable and the optimization convex in nature; however, they render the end result less useful for practical implementation. In contrast to a dense HetNet as the main motivation for SON functions, some works exist, wherein the solution is proposed and simulated mainly for macrocell scenarios; i.e., large CIOs and Tx power disparities between small cells (SCs) and macro cells are not considered. While these approaches may work for current network deployment, they will not be useful for the dense HetNet architecture envisioned for 5G.

1.2 Research Objectives

In light of the discussion in section 1.1, the research presented in this dissertation provides answers to the following questions:

1. How does one perform self-optimization with low time complexity through just top-level policy input that prioritizes otherwise conflicting key performance indicators?

2. A user mobility pattern is known to have a high predictability component. Is there a way in which this predictability can be exploited to predict the future location of users and thus set foundations for proactive self-organizing network functions?

3. If small cell densification is evident for future network growth, how can we proactively reduce the high aggregated network energy that "always ON" small cells are bound to consume in an ultra-dense network, while meeting ambitious 5G quality of experience requirements?

4. Imbalanced loads among small and macro cells and poor resource utilization as a consequence form a paramount challenge that hinders wide-scale ultra-dense network deployments. How can one perform load balancing in a proactive way to meet the extremely low latency and high quality of experience expected from 5G and beyond?

5. Anticipated high operational complexity and cell density in 5G indicates that an ultra-dense 5G network is bound to face significantly higher cell outage rates than legacy networks. How should the susceptibility of an ultra-dense, extremely complex 5G network to a potentially high cell outage rate be managed?

This dissertation addresses the aforementioned research questions. Analytical models are developed, and 3GPP-compliant rigorous simulation studies are carried out to find and validate the answers to the above questions. The key contributions of the dissertation are outlined in the following section.

1.3 Contributions

The contributions of this dissertation can be summarized as follows:

- This dissertation contributes by presenting a generic low-complexity cellular system optimization framework to provide the agile, on-line, multi-objective optimization of potential network topology configurations (NTCs) that can judiciously strike the intended balance between the various conflicting goals, such as capacity, QoS, and power consumption, while taking into account an operator's policy. This framework quantifies, analyzes, and optimizes the three major KPIs used for the holistic optimization of SON-enabled heterogeneous cellular systems, namely, capacity, service area fairness (SAF) and power consumption. This framework can model the KPIs of interest as functions of a comprehensive set of optimization parameters such as the spectrum reuse factor, the number of sectors per site, the number of SCs per site, adaptive coding, and modulation. The metrics derived can be quickly evaluated semi-analytically and thus facilitate a solution to the multi-objective, holistic optimization problem that is otherwise tackled using black box-type complex, dynamic simulation models. Using the proposed performance characterization framework (PCF), we also evaluate and compare 26 different network topologies and quantify their relative gains. We analyze the respective trade-offs offered by each NTC in terms of capacity, SAF and power consumption. Our results also demonstrate that contrary to common notion, NTCs with the highest spectrum efficiency are not necessarily those that resort to full frequency reuse. The insights provided by the proposed framework can help to address new requirements from future heterogeneous cellular networks. Building on these insights, we propose a heuristic algorithm named "classify parameters, prioritize objectives, and solve subproblems (CPS)" for holistic optimization. Through a case study, we demonstrate, how the PCF and CPS together can be used for a wide range of cellular optimization scenarios with low complexity.

- Next, we develop and analyze a semi-Markov model-based spatio-temporal mobility prediction framework. Our proposed mobility prediction model overcomes the limitation of conventional discrete-time Markov chain (DTMC)-based prediction models that fail to incorporate the time dimension, i.e., "Time of next Hand Over (HO)." Next, we propose a novel method to map the next cell's spatiotemporal HO information to the estimated future location coordinates based on the idea of Landmarks. This novel method further increases the spatial resolution of the future location estimation without requiring an increase in the number of states for the semi-Markov model. The accuracy of the proposed model is quantified through experimental evaluations coupled with extensive Monte Carlo simulations.

- Another contribution is AURORA wherein, based on the intelligence gained from the mobility model, i.e., future cell loads, a proactive energy-saving (ES) optimization problem is formulated to minimize the energy consumption by switching OFF underutilized SCs. In addition to proactiveness, another key novelty of the proposed ES scheme is that it leverages CIOs as optimization variables for balancing the load between cells while deciding which cells to switch ON/OFF. In this way, an additional UDN-specific mechanism is exploited to ensure the QoS while maximizing ES. Although the formulated problem is non-convex, large-scale, combinatorial, and NP-hard, our results indicate that the structure of the problems allows heuristics such as genetic programming to find useful solutions with high ES yield. The ahead-of-time estimation of cell loads allows ample time for such heuristics to converge without jeopardizing the QoS. We conduct multi-tier, system-level 3GPP-compliant rigorous simulations for a comprehensive performance analysis of the proposed AURORA. The prediction accuracy of the semi-Markov-based mobility prediction model has been quantified using the realistic SLAW mo-

bility model in a HetNets environment. The average location estimation error was found to be around 28 meters, while relying only on one piece of information that is already available in the network, namely, HO trace. We also analyze the impact of cell load thresholds on ES gains and the QoS (percentage of satisfied users) for proactive ES optimization. The results of this analysis provide actionable insights for determining cell load thresholds that can judiciously strike the intended balance between the conflicting goals of ES and QoS. We perform a comparative analysis of the proposed solution in low and high traffic demand scenarios, with the latter comprising all video users, against several benchmarks, including industrial practices, i.e., All-ON SCs without and with fixed CIOs. AURORA achieves significant gains in the total network energy reduction for low and high traffic demand scenarios by putting under-utilized SCs in sleep mode with a negligible number of unsatisfied users. We also investigate a deep neural network (DNN)-based mobility prediction model to test the sensitivity of AURORA to mobility prediction model and its accuracy. DNN-based mobility prediction model offers slightly higher prediction accuracy and hence better performance gain in AURORA, compared to semi-Markov but at the cost of substantial increased complexity and training time. Moreover, we compare AURORA with a near-optimal performance bound that is achievable when future network load conditions can be estimated with 100% accuracy. This comparison demonstrates that AURORA is reasonably resilient to location estimation inaccuracies.

- The next contribution is the OPERA framework that can leverage the knowledge gained from mobility/hand-off patterns for coping with load imbalance challenges in 5G and beyond. The proactiveness of OPERA stems from its novel capability that instead of passively waiting for congestion indicators to be observed and then reacting to them, OPERA predicts future cell loads us-

ing readily available data streams such as past HO traces, and then proactively optimizes key network parameters that affect cell load and network capacity namely azimuths, beam widths, Tx power and CIOs to preempt congestion before it happens. Although the resulted problem is NP-hard, the ahead of time estimation of cell loads allows ample time for a dexterous combination heuristics such as genetic programming and pattern search to find solutions with high gain. We use extensive system level simulations to evaluate OPERA and compare its performance against three different benchmarks: (i) real network deployments settings taken from an LTE operator, (ii) recently proposed LB scheme in literature as representative of state-of-the-art reactive schemes, and (iii) upper performance bound where user future location is assumed to be known with 100% accuracy. Realistic SLAW model based mobility traces are used in the performance analysis. Results show that compared to benchmarks, OPERA can yield significant gain in terms of fairness in load distribution and percentage of satisfied users. Superior performance of OPERA on several fronts compared to current schemes stems from its following features: 1) It preempts congestion instead of reacting to it; 2) it actuates more parameters than any current LB schemes thereby increasing system level capacity instead of just shifting it among cells; 3) while performing LB, OPERA simultaneously maximizes residual capacity while incorporating throughput and coverage constraints; 4) it incorporates a load aware association strategy for ensuring conflict free operation of LB and CCO SON functions.

- Lastly, this dissertation contributes by presenting a stochastic analytical model to analyze and predict the arrival of faults on the reliability behavior of a cellular network. Assuming exponential distributions for failures and recovery, a reliability model is developed using the CTMC process. The proposed model, unlike previous studies on network reliability, is not limited to structural as-

pects of BSs. It also takes into account diverse potential fault scenarios, and it is capable of predicting both the expected time of the first occurrence of the fault and the long-term reliability behavior of the BS. This model can adapt itself dynamically by learning from a past database of network failures. Three different scenarios have been analyzed in terms of transient analysis, occupancy time, first passage time, and the steady-state distribution. As per the numerical results, the mean arrival rate of trivial failures has a profound effect on the reliability behavior of the cellular network. Another key finding is that a substantial gain in network reliability can be achieved by reducing a BS's fault detection and recovery time; this strongly advocates the need for agile self-healing SON functions.

1.4 Dissemination and Publications

Throughout the course of preparation for this dissertation, several dissemination activities were carried out. These activities have resulted in the following presentations and (accepted or pending) peer reviewed articles.

Awards:

A1. Awarded Gallogly College of Engineering Dissertation Excellence Award by University of Oklahoma.

A2. Winner of the IEEE Young Professional Green ICT Idea Competition 2017. The core idea of AURORA framework has won IEEE GREEN ICT Best Solution Award Competition 2017. Received the award at IEEE Greening through ICT Summit held in Paris, France on 3rd October 2017.

A3. Winner of a nationally competitive travel grant for participation in IEEE ComSoc Summer School 2017 held in New Mexico, USA.

A4. Awarded twice for Best Research Presentation at TCOM Research Meeting, University of Oklahoma. Received an honorable mention for having highest score in the past decade.

Patents:

P1. A. Imran, A. Asghar and H. Farooq, "Method for enhancement of capacity and user Quality of Service in Mobile Cellular Networks", provisional patent application number: 62681320 filed June 06, 2018.

P2. A. Imran, H. Farooq and A. Asghar, "Method and apparatus for proactive self-optimization using data about network user behavior, mobility and measurements", 2018 (pending provisional patent application).

Book Chapters:

B1. H. Farooq, A. Imran and M. S. Parwez, "Continuous Time Markov Chain Based Reliability Analysis for Future Cellular Networks", Big Data Applications in the Telecommunications Industry, Ye Ouyang and Mantian Hu, IGI Global, pp 119-136, 2016.

B2. M. S. Parwez, H. Farooq, A. Imran and H. Refai, "Spectral Efficiency Optimization Using Clustering and Dynamic Beam Steering in Self Organizing Cellular Networks", Big Data Applications in the Telecommunications Industry, Ye Ouyang and Mantian Hu, IGI Global, pp 137-155, 2016.

Journals:

J1. H. Farooq, A. Asghar and A. Imran, "Mobility Prediction based Automated Proactive Energy Saving (AURORA) Framework for Emerging Ultra-Dense Networks", IEEE Transactions on Green Communications and Networking, 2018, DOI: 10.1109/TGCN.2018.2858011.

J2. A. Asghar, H. Farooq and A. Imran, "Concurrent Optimization of Coverage, Capacity, and Load Balance in HetNets Through Soft and Hard Cell Association Parameters", IEEE Transactions on Vehicular Technology, vol. 67, no. 9, pp. 8781-8795, Sept. 2018.

J3. A. Asghar, H. Farooq and A. Imran, "Self-Healing in Emerging Cellular Networks: Review, Challenges, and Research Directions", IEEE Communications Surveys & Tutorials, vol. 20, no. 3, pp. 1682-1709, 2018.

J4. A. Taufique, A. Mohamed, H. Farooq, A. Imran, and R. Tafazolli, "Analytical modelling for mobility signaling in ultra-dense hetnets", IEEE Transactions on Vehicular Technology, 2018, DOI: 10.1109/TVT.2018.2846655.

J5. A. Zoha, A. Saeed, H. Farooq, A. Rizwan, A. Imran, and M. A. Imran, "Leveraging intelligence from network CDR data for interference aware energy consumption minimization", IEEE Transactions on Mobile Computing, vol. 17, no. 7, pp. 1569-1582, July 2018.

J6. A. Said, S. W. H. Shah, H. Farooq, A. N. Mian, A. Imran and J. Crowcroft, "Proactive Caching at the Edge Leveraging Influential User Detection in Cellular D2D Networks", Future Internet, vol. 10, no. 10, p. 93, Sep. 2018.

J7. S. Bassoy, H. Farooq, M. A. Imran and A. Imran, "Coordinated Multi-Point Clustering Schemes: A Survey", IEEE Communications Surveys & Tutorials, vol. 19, no. 2, pp. 743-764, 2017.

J8. H. Farooq and A. Imran, "Spatiotemporal Mobility Prediction in Proactive Self-Organizing Cellular Networks", IEEE Communications Letters, vol. 21, no. 2, pp. 370-373, Feb. 2017.

J9. H. Farooq, A. Imran and A. Abu-Dayya, "A Multi-objective Performance Modeling Framework for enabling Self-Optimization of Cellular Network Topol-

ogy and Configurations", Transactions on Emerging Telecommunications Technologies, vol. 27, pp. 1000-1015, 2016.

J10. A. Asghar, H. Farooq and A. Imran, "Entropy Field Decomposition Based Time and Space Aware Outage Detection Solution for Ultra-Dense Millimeter Wave Heterogeneous Networks", IEEE/ACM Transactions on Networking (under review).

J11. S. M. A. Zaidi, A. Taufique, H. Farooq, and A. Imran, "Mobility Management in 5G and Beyond: A Survey Outlook", IEEE Communications Surveys & Tutorials (under review).

J12. H. Farooq, A. Asghar and A. Imran, "Mobility Prediction based Proactive Dynamic Network Orchestration for Load balancing with QoS Constraint (OPERA)", IEEE/ACM Transactions on Networking (under review).

J13. H. Farooq, A. Taufique and A. Imran, "Challenges in 5G Networks Mobility Management: How to change Mobility from bane to blessing?", IEEE Networks (under review).

Conferences:

C1. H. Farooq, A. Asghar, and A. Imran, "Mobility Prediction empowered Proactive Energy Saving Framework for 5G Ultra-Dense HetNets", accepted in Proc. IEEE GLOBECOM'18, Abu Dhabi, United Arab Emirates, 2018.

C2. A. Asghar, H. Farooq, and A. Imran, "Concurrent CCO and LB Optimization in Emerging HetNets: A Novel Solution and Comparative Analysis", accepted in Proc. IEEE PIMRC'18, Bologna, Italy, September 2018.

C3. Hughes, S. Bothe, H. Farooq and A. Imran, "Generative Adversarial Learning for Machine Learning empowered Self Organizing 5G Networks", accepted in

Proc. Workshop on Computing, Networking and Communications (CNC) associated with International Conference on Computing, Networking and Communication (ICNC) Hawaii, 2019.

C4. M. N. Rafiq, H. Farooq, A. Zoha and A. Imran, "Can Temperature Be Used as a Predictor of Data Traffic: A Real Network Big Data Analysis", accepted in Proc. 5th IEEE/ACM International Conference on Big Data Computing, Applications and Technologies (BDCAT), Zurich, 2018.

C5. A. Asghar, H. Farooq, and A. Imran, "A Novel Load-Aware Cell Association for Simultaneous Network Capacity and User QoS Optimization in Emerging HetNets", in Proc. IEEE PIMRC'17, pp. 1-7, 2017.

C6. Y. Kumar, H. Farooq and A. Imran. "Fault prediction and reliability analysis in a real cellular network", in Proc. 13th International Wireless Communications and Mobile Computing Conference (IWCMC), Valencia, Spain, pp. 1090, 2017.

C7. H. Farooq, A. Imran and M. S. Parwez, "Continuous Time Markov Chain Based Reliability Analysis for Future Cellular Networks", in Proc. IEEE GLOBECOM'15, pp. 1-6, San Diego, CA, 2015.

C8. M. S. Parwez, H. Farooq, A. Imran and H. Refai, "Spectral Efficiency Optimization Using Clustering and Dynamic Beam Steering in Self Organizing Cellular Networks", in Proc. IEEE GLOBECOM'15, pp. 1-7, San Diego, CA, 2015.

C9. H. Gebrie, H. Farooq, and A. Imran, "Comparative Performance Analysis of Machine Learning Techniques for Mobility Prediction in Cellular Networks", (under review).

C10. A. Asghar, H. Farooq and A. Imran, "Outage Detection for Millimeter Wave Ultra-Dense HetNets in High Fading Environments", (under review).

C11. H. Farooq, A. Asghar, and A. Imran, "Mobility Prediction based Proactive Dynamic Network Orchestration for QoS aware Load balancing", (under review).

1.5 Organization

The dissertation is structured as follows. Chapter 2 presents the background on SONs; it highlights 3GPP standardized SON use cases. Chapter 3 presents a generic low-complexity optimization framework coupled with a heuristic algorithm to provide agile, on-line, multi-objective optimization of future MCN through just top-level policy input that prioritizes otherwise conflicting KPIs such as capacity, QoS, and power consumption. Chapter 4 presents the semi-Markov renewal process-based mobility prediction model that predicts the future location of users and hence transforms mobility from a challenge into an advantage. This predicted load distribution of the cells is then used to formulate a novel (i) energy saving optimization problem (AURORA) in chapter 5 that proactively schedules SC sleep cycles and (ii) proactive LB optimization problem (OPERA) in chapter 6. The two contributions on proactive SON (P-SON) leverage the fact that ahead-of-time estimation of cell loads allows ample time for heuristics such as genetic programming to find solutions with high ES and LB yields. Chapter 7 presents a stochastic reliability analytical model to predict the expected time of the first occurrence of the fault and the long-term reliability behavior of the BS. Finally, chapter 8 discusses the conclusions and future work, and it thus concludes the dissertation.

CHAPTER 2

Background

> It is not the strongest of the species that survives, nor the most intelligent that survives. It is the one that is the most adaptable to change.
>
> Charles Darwin

This chapter is devoted to a SON as a feature of 3GPP LTE systems. It briefly presents the basic concepts, and it provides a high-level introduction to the structure of 3GPP SON use cases. It goes through self-optimization and healing functionalities, introducing the use cases that have been defined for each of them and mainly talking about how they have been addressed in 3GPP. It does not focus on the general literature on SON, as this has already been reviewed in other works [4]. Furthermore, it discusses the self-coordination problem that may exist between the parallel execution of multiple SON functions. It describes the MDT functionality, which is introduced to collect useful data for analysis from UE measurements, in order to improve coverage and capacity issues, and verify the QoS among other things. Finally, it highlights big data sources in cellular networks that need to be exploited by SONs to become 5G viable.

2.1 Introduction to SONs

As the spectral efficiency per link for LTE is approaching the theoretical Shannon limit, it is envisaged that the network densification by SCs is among the most

promising solutions for realizing ambitious goals of infinite capacity and zero latency provision in future 5G networks. It is not difficult to prognosticate that the complexity of operation and OPEX of the resultant ultra-dense network is bound to become the largest challenge in 5G and beyond. To cope with a much punier version of this challenge in 4G, research on SONs has already begun in recent years, mainly to reduce the operational cost of manual labor. However, since the bulk of target capacity gain in 5G must come from network densification, even the technical feasibility of 5G hinges on its SON capabilities, in addition to financial viability. This paradigm of SON has recently been heavily investigated to automate cellular system management and maintenance tasks [4]. The main objective of SON is to reduce cost, i.e., CAPEX and OPEX, by minimizing human involvement, while enhancing network performance, in terms of network capacity, coverage, and QoE. The main motivation behind the increasing interest in the introduction of SON from standardization bodies and operators is twofold. On the one hand, from a technical perspective, the complexity and large scale of future radio access technologies imposes significant operational challenges due to the multitude of tuneable parameters and the intricate dependencies between them. With each successive generation of cellular networks, the complexity of BSs has continued to increase; i.e., typical 2G, 3G, and 4G cells have roughly 500, 1000, and 1500 parameters respectively to optimally configure and maintain. Without intervening measures, the same complexity growth trend is expected for 5G [5]. Traditional network management using classic manual and field trial design approaches are hence no longer viable. The overall idea of SON is to integrate network planning, configuration, and optimization into a single cognitive, automated process requiring minimal human intervention. 3GPP has defined the number of SON use cases first introduced in Release 8 and expanding to subsequent releases. These meaningful SON use cases can be classified as follows according to the life cycle phases of a mobile network

(planning, deployment, optimization, and maintenance).

1. **Self-Configuration**–automation of network configuration and planning of newly deployed BSs.

2. **Self-Optimization**–automation of the tuning of a deployed mobile cellular infrastructure's COP for obtaining the best network configuration and performance over time.

3. **Self-Healing**–automation in detection, diagnosis, compensation, and recovery from the failures of a deployed mobile cellular infrastructure.

Self-optimization and Self-healing are two of the most important functionalities in SONs because they ensure that the network operates at its best level of efficiency once the BSs have been deployed. Since this PhD dissertation is focused on self-optimization and self-healing, the following section further describes these functions.

2.2 Self-Optimization

Self-optimization algorithms aim to optimize ongoing services in the network based on network measurements. These algorithms monitor network performance data and perform optimization changes in the network in open and/or closed loops, aiming to reduce OPEX costs as well as improve network performance in terms of network spectral efficiency, EE, network capacity, and the overall QoS. Self-optimization is a core part of LTE/LTE-Advanced standardization, and commercialized algorithms are already deployed in current LTE networks. Specific use cases are described as follows:

1. **Mobility load balancing (MLB):** Load balancing aims to balance the load between the available cells in a certain geographical area. The goal of LB

in general is to move traffic from high loaded cells to less loaded neighbors as far as the interference and coverage situation allows. In this way, better utilization of cell capacity and larger UE throughputs can be reached. The implementation of this function is generally distributed and supported by the load estimation and resource status exchange procedure.

2. **Mobility robustness optimization (MRO):** The MRO is a SON function designed to guarantee proper mobility, i.e., proper handover in connected mode and cell re-selection in idle mode between cells of the same and different radio access technologies (RATs). Common targets include reduced call drops, the minimization of radio link failures (RLFs), and a decreased number of ping pongs. The messages containing useful information are as follows: the S1AP handover request or X2AP handover request, the handover report, the RLF indication/report. Mobility robustness optimization operates over connected mode and idle mode parameters. In connected mode, it tunes meaningful handover trigger parameters, such as the event A3 offset (when referring to intra-RAT, intra-carrier handovers), the time-to-trigger, or the Layer 1 and Layer 3 filter coefficients. In idle mode, it tunes the offset values, such as the Qoffset for the intra-RAT, intra-carrier case.

3. **Inter-cell interference coordination (ICIC):** The ICIC minimizes interference among cells sharing the spectrum. Its working hinges on the coordination of physical resources between co-channel neighboring cells to reduce interference from one cell to another. ICIC can be static, semi-static, or dynamic, wherein ICIC relies on frequent adjustments of parameters, supported by signaling among cells over an X2 interface.

4. **Random access channel (RACH):** The RACH finds the best trade-off between the performance of the random access and the resources that have

to be sacrificed for it based on UE feedback and knowledge of its neighboring eNBs RACH configuration. Random access channel optimization can be done by adjusting the power control parameter or changing the preamble format to reach the set target access delay.

5. **Coverage and capacity optimization:** The main objective of the CCO use case is to provide sufficient coverage and capacity in the whole network area with minimal radio resources. This CCO use case can be further divided into two sub-objectives:(i) maximizing the relative coverage in the area so that continuous coverage would be achieved, where the relative coverage can be defined as the probability that the received signal quality is better than the minimum required received signal quality, and (ii) providing a sufficient received quality in terms of an achievable bit rate over the entire area.

6. **Energy saving:** The deployment of ES is motivated by the goals to reduce CO_2 footprint and to optimize the costs. The radio access network (RAN), and particularly the radio BSs, have been identified as having the highest share of mobile networks' overall energy consumption and hence the largest potential for ES measures. Energy saving aims to provide desired QoE to end users with minimal impact on the environment. Energy saving achieves its objective by temporarily switching OFF unused capacity when not needed.

2.3 Self-Healing

Self-healing comes into play during the maintenance phase of a cellular network. Wireless cellular systems are prone to faults and failures, and the most critical domain for fault management is the RAN. A complete (or partial) cell outage is a scenario when either a BS's hardware and/or software malfunctions or when one or more cell parameters become misconfigured during network operations. Partial

outage refers to scenarios when the cell continues to operate but its performance degrades below its typical level. The cell outage rate in a network is intrinsically proportional to the number of cells and number of components and parameters per cell. Forthcoming cellular networks are susceptible to even higher cell outage rates caused by the misconfiguration of parameters due to potential conflicts between multiple SON functions. Self-healing performs an automatic adjustment of network parameters and algorithms in surrounding cells to compensate the outage users until the problem is solved. Once the actual failure has been repaired, all parameters are restored to their original settings. Self-healing solutions are broadly classified as follows:

1. **Cell outage detection**–these solutions aim to detect a cell outage through the monitoring of performance indicators, which are compared against thresholds and profiles.

2. **Cell outage compensation**–this use case aims to provide service to users who were previously served by a cell in outage by appropriately adjusting suitable radio parameters, such as the pilot power and the antenna parameters of the surrounding cells.

A significant number of self-optimization and self-healing solutions for 4G have recently been proposed by the research community. However, there are dramatic differences between current cellular systems and the 5G cellular networks that call for a paradigm shift in SON research to enable a commercially and technically feasible 5G network.

2.4 Self Organizing Network Evolution in 3GPP

3GPP Release 8 includes SON functionalities relating to the initial equipment installation and integration [6]. The SON functionality developed in Release 9 focuses on the optimization of deployed LTE networks, and Release 10 introduces SON functions to offset interoperability issues in HetNets and includes NGMNs recommendations. Release 11 SON functions are related to the self-management of heterogeneous networks, while Release 12 focus on enhancing the performance of the centralized CCO functions. Release 13 studies schemes for enhancements of the OAM aspects of distributed LB as well as enhanced centralized CCO. Finally, Release 14 focuses on meeting the ambitious 5G requirements in terms of zero latency, fair co-existence in an un-licensed spectrum, EE, support for carrier aggregation, and SON support for active antennas.

2.5 Self-Organizing Network Architecture

Depending on the location of SON functions or where they are executed, three possible architectures have been considered for a SON defined as (i) a centralized SON (C-SON), where SON algorithms reside in the network management system or in OMC; (ii) D-SON, wherein the SON functions are distributed across the edges of the network; and (iii) a hybrid SON, with SON algorithms located at different levels. C-SON style algorithms take input from all nodes in the network and have a global picture of the overall network at the cost of low agility which may be pronounced in the emerging world of SCs that experience highly transitory traffic loads. On the other hand, D-SON functionalities have high agility, which enables the network to adapt to local changes more rapidly at the cost of higher vulnerability to network instabilities that maybe caused, e.g., by the concurrent operation of SON functions with conflicting objectives/parameters/KPIs. Therefore, it is crucial to select the

SON function execution location and thus the SON architecture mainly on a per use case basis. From a design perspective, SON functions must have the following capabilities:

- Autonomy–the network must be able to operate autonomously; i.e., SON functions must be independent of human input during operation.
- Scalability–the functions deployed within SONs must be scalable enough to be able to run within the limited computational capabilities afforded by the network nodes; i.e., any SON functions deployed in the network must be scalable in terms of both time and space.
- Adaptability–the network must be able to adapt to outside influences and internal failures.
- Agility–SON functions should have low time complexity and thus high agility to meet the zero latency requirements of 5G.
- Cognition–SON networks must be intelligent; i.e., they must be able to learn from the information generated by users and network entities to become completely independent in terms of adapting network parameters based on the primary goals of the mobile network operators (MNOs).

2.6 Self-Coordination

Self-organizing network algorithms, being use-case centric, are often designed as stand-alone functionalities. As a result, the concurrent operation of multiple independent SON functions is prone to unhealthy conflict when implemented together in a network. As identified in [3], in an uncoordinated SON, a variety of conflicts may occur when 1) two or more SON functions try to modify the same network

configuration parameter; 2) a SON function is triggered by an input parameter whose value is dependent on some other network parameters; 3) there is a change in network conditions by the impromptu addition or removal of relay, eNB, or Home eNB; 4) different SON function actions try to alter the same KPI of a cell while adjusting different network configuration parameters; 5) a SON function computes new parameter configuration values based on outdated measurements; and 6) there is a logical dependency between the objectives of SON functions. The uncoordinated working of multiple SON functions can thus be subject to a large number of potential conflicts, which can actually degrade a network's performance instead of improving it. For example, CCO may try to improve coverage by increasing Tx power which may force a large number of users to jump into its coverage, thereby conflicting with the LB SON objective. Therefore, it is deemed necessary to consider SON conflicts when devising SON functions.

2.7 Minimization of Drive Tests

The true potential of SONs is contingent on the timely availability of network measurements. Therefore, a key enabler for SON functions are MDT Reports, which were standardized in 3GPP Release 10. Minimization of drive test reports enables the network to instruct UE's to log network measurements–such as the reference signal received power (RSRP) and reference signal received quality (RSRQ) of serving and neighboring cells–and send them back to the core through radio resource control (RRC) signaling messages, thereby avoiding manual and time consuming physical drive tests. The MDT scheme supports two reporting configurations: (i) non-real time or immediate reporting wherein when the preconfigured triggers are met, the UE immediately reports the measured radio conditions, and (ii) logged mode wherein UE stores measurements and reports them when the periodic timer expires. When eNB requests UE for MDT reports through the *UEInformationRe-*

quest RRC signaling message, UE responds by embedding the desired measurement results in the *UEInformationResponse* RRC message and sending it back to the network. Moreover, the measurement reports are tagged with the location information of reporting UEs for facilitating SON algorithms. Based on the intelligence extracted from the measurements received from UEs, the SON engine can initiate appropriate SON functions to achieve optimum network performance that aligns with the business's tailored objectives prescribed by the operator. This MDT feature is effectively a pre-requisite for enabling the range of SON functions that either have already been standardized by 3GPP or are being considered for the future evolution of SON.

2.8 Big Data Sources for SON

Mobile networks routinely produce massive amounts of control, signaling, and contextual data during the day-to-day operation of cellular networks–also referred to as big data. When exploited, this big data can be an enabler for a paradigm shift in SONs to meet the ambitious QoS requirements of 5G. The potential constituents of big data in cellular networks are as follows [5]:

- a) Subscriber-level data. These data comprise of KPIs obtained from a voice or a data session initiated by the subscriber to provide an indication of the accessibility, retainability, and integrity performances of the network. Several metrics including blocked call rates, access failure rates, setup times, the success rate, and hand-over failure rates, project the accessibility of the network. Dropped call rates, completion times, the packet data protocol context, and success rate together define the retainability of the network. Metrics such as speech and data streaming quality, throughput, packet jitter, and delay offer insight into a user's perceived QoE.

- b) Cell-level data. This refers to the measurements that are reported by a BS and all users within the coverage of that BS. Examples of useful cell-level data streams are measurements reporting the uplink noise floor in terms of reference interference power, channel-based power information, physical resource block (PRB) usage per cell, the number of active users per cell, and MDT measurements. Minimization of drive test reports consist of the RSRP and RSRQ values of the serving and neighboring cells reported by the users to their serving BSs.

- c) Core network-level data. Core network data include signaling information, historical alarm logs, equipment configuration lists, and service and resource utilization accounting records (call data records - [CDRs] and extended data records [XDRs]) as well as the aggregate statistics of network performance metrics.

- d) Miscellaneous data. These data consist of the structured information already stored in the separate databases, including customer relationship management, customer complaint center, and spectrum utility maps. They also include un-structured information such as social media feeds, specific application usage patterns, and data from smart-phone built-in sensors and applications.

2.9 Conclusion

In this chapter, we described the field of SONs by providing an overview of the 3GPP standardized SON use cases in the domains of configuration, optimization, and healing and with regard to the need for self-coordination. Then, we highlighted big data sources in cellular networks that, when exploited, can make SONs viable for 5G networks.

CHAPTER 3

A Multi-Objective Performance Modeling Framework for Enabling the Self-Optimization of Cellular Network Topology and Configurations

Strength without agility is a mere mass.

Fernando Pessoa

Cellular system optimization (CSO), a cornerstone of the cellular systems paradigm, requires a new focus shift because of the emergence of a plethora of new features shaping the cellular landscape. These features include SONs with added flavours of heterogeneity of cell sizes and BS types, adaptive antenna radiation patterns, EE, spatial homogeneity of service levels, and a focus shift from coverage to capacity. Moreover, to effectively tackle the spatiotemporal dynamics of network conditions, a generic low-complexity framework to quantify the key facets of performance–namely, the capacity, QoS and power consumption–of the various NTCs, is needed to enable SONs driving the cellular system optimization on the fly. In this chapter, we address this problem and presenting a PCF that quantifies the multiple performance aspects of a given heterogeneous NTC through a unified set of metrics that are derived as a function of key optimization parameters. We then leverage this framework to present a cross comparison of a wide range of potential NTCs. Moreover, we propose a low-complexity heuristic approach for the holistic optimization of future heterogeneous cellular systems for joint optimality in multiple desired performance indicators. The PCF also provides quantitative insights into the new tradeoffs involved in the optimization of emerging heterogeneous

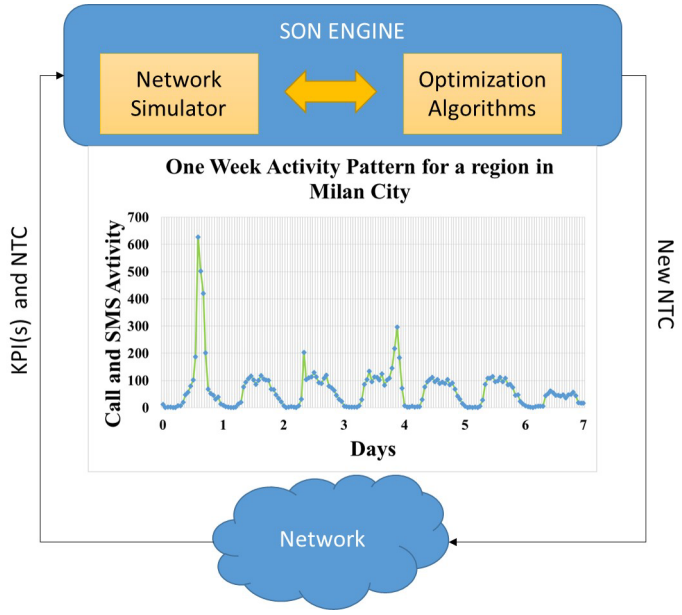


Fig. 3.1: A SON engine has to cope with frequent activity variations observed in a real cellular network

networks, and it can pave the way for much needed further research in this area.

3.1 Introduction

The paradigm of SONs has recently been heavily studied to automate cellular system management and maintenance tasks [4, 7, 8]. This SON capability allows a cellular network to monitor KPIs and optimize network parameters to adapt itself to the spatiotemporal dynamics of network conditions. These dynamics include a change in traffic patterns over the course of a day, the relocation of hot spots, and cell outages. For example, Fig. 3.1 depicts a SON-enabled cellular network and the variations in traffic patterns observed in a real cellular network's data. These variations exhibited through the KPIs and that can further be estimated using MDT reports [9], prompt the SON engine to test each of the possible NTCs in a static or dynamic simulator to devise a new NTC that meets specific objectives such as spectral efficiency,

EE, QoS, or a combination of these. Leveraging the modern capability of turning BSs on and off and smart antenna radiation patterns, the SON engine can adapt a projected number of sectors, frequency usage, and the number of SCs on the fly to achieve the desired objectives. In a real network, there are hundreds of possible network parameters configurations (i.e., large search space) characterized by the types of BSs, the number of sectors per site, the number of SCs per site, and the frequency reuse, in addition to other configuration parameters, including, locations, tilts, azimuths, and heights. In this ever-changing traffic landscape of the cellular environment, by the time a SON engine comes up with an optimum network configuration, the scenario might have already changed and the NTC becomes outdated. This calls for a low-complexity performance estimation and then optimization techniques to cope with the spatio-temporal dynamics of the cellular environment in an agile fashion. Increasing the scarcity of the spectrum for LTE is pushing for more aggressive frequency reuse, leading to new kinds of spectrum reuse, e.g., intra-site spectrum reuse [10, 11] that has to be incorporated into SON optimization objectives. Also, fueled by the performance criteria set forth by 3GPP, where the spatial fairness of the data rate received by the cell edge and cell center users is being assigned increasing importance [12], the QoS metric can no longer be neglected. Similarly, in the wake of the rising cost of energy and environmental concerns, power consumption has also become an important metric [13].

The ambitious goals of zero latency [5] in envisioned future cellular systems (CSs) require a low-complexity CSO framework to provide an agile, on-line, multi-objective optimization of potential NTCs that can judiciously strike the intended balance between the various conflicting goals, such as capacity, QoS and power consumption while taking into account an operator's policy. The

need for the potential search of NTCs, though well-conceived in [14, 15, 16, 17], is not fully addressed yet, particularly in the context of SCs enhanced CS (SC-CS) HetNets. Another challenge in enabling and evaluating many of the SON use cases in HetNets is the lack of a unified performance quantification framework that can quantify cellular system performance in terms of the aforementioned KPIs. This chapter addresses that need by presenting and analyzing a holistic framework to quantify the three KPIs, namely, capacity, QoS, and power consumption. This framework can act as a key enabler for a number of SON use cases such as CCO, ICIC, EE and LB.

3.1.1 Prior Works

For cellular networks, most of the prior research studies have focussed on the optimization of network parameters [18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37], using different definitions of a given KPI, e.g., coverage and capacity [19, 20, 21, 31, 32, 33, 18], QoS [22, 23], cost efficiency [24], or EE [25, 26, 34, 35], to optimize a single network parameter, e.g., the BS location [26, 27, 28, 38], or few other parameters such as antenna tilts [31], sectorization [29, 36], and frequency reuse [30, 37]. Moreover, these KPI metrics, which are to be used by the SON engine, should be able to quantify the long-term average performance of a cellular system by incorporating its dependencies on NTC parameters, while generalizing or averaging out the short-term dynamics of the cellular eco-system. An additional requirement is that the metrics should be evaluable by the SON engine without resorting to complex dynamic simulators. More precisely, to the best of our knowledge, no previous work has provided a framework to enable a cross-comparison of potential NTCs, in terms of capacity, QoS and power consumption simultaneously, while taking into account key deployment factors such as the number of

sectors per site, the number of SCs per site, and different variants of intra-site frequency reuse that the emerging CSs can avail. Furthermore, the trade-off between the capacity and spatial fairness of a service level in the coverage area is relatively overlooked. In view of the increasing emphasis by 3GPP on better cell edge throughput rates and better spatial fairness of achievable data rates [23], we also use the proposed PCF to investigate this under-explored but important trade-off that various NTCs offer. The presented analysis can be leveraged to design NTCs that can strike an operator-intended precise balance between the capacity and spatial fairness while simultaneously taking into account the power consumption aspect of the given NTC.

Additionally, since the optimization of network parameters is an NP-hard problem, prior works in literature have generally addressed it using meta-heuristics such as simulated annealing [39, 40, 38], particle swarm [41], genetic algorithms (GAs) [42, 33], Taguchi's method [43], or ant colony optimization [29] to obtain near optimal solutions for a selected set of few parameters. The basic methodology that is generally followed in these works involves a detailed dynamic simulation model that acts as a black box between the KPI and the potential parameters of a given NTC. The SON engine's use of dynamic simulation-based models is not only time consuming, but it also provides little insight into system behavior. In contrast, our approach builds on a mathematical model to couple the KPIs with the extensive set of NTC parameters and thus helps to obtain better insights into system behavior. The resultant PCF makes the holistic cross-comparison of various potential solutions to the CSO problem easier.

3.2 Background and System Model

3.2.1 Cellular System Optimization Objectives and Proposed Solution Approach

For emerging CSs, the optimization problem has multiple target objectives, such as the maximization of capacity, coverage, fairness of service in the coverage area, spectral efficiency, spectrum reuse efficiency, throughput, minimization of cost, energy consumption and/or outage. However, all these objectives can be boiled down to three main categories of performance measures:

1. Capacity oriented performance measures–these include cellular capacity, spectral efficiency, spectrum reuse efficiency, throughput, or goodput.
2. Quality-of-Service oriented performance measures–rate fairness and outage are typical QoS measures.
3. Cost oriented performance measures–the total cost of ownership of a cellular system over its life has three further major factors:
   (a) Capital cost–cost of hardware, cost of software, and deployment labor cost.
   (b) Maintenance cost–cost of labor required for operation, optimization, and maintenance of sites and the switching network.
   (c) Power consumption–power consumed to keep the cellular system running is increasingly becoming a significant factor of operational cost.

In this chapter, we derive the PCF to quantify each of the three listed aspects of a cellular system's performance as a function of NTC parameters. Under cost-oriented performance, we only focus on the power consumption, as it has recently become highly important, particularly due to rising costs of energy

and concern for CO2 emissions. For the treatment of the other two cost factors, interested readers are referred to previous works in [14], which deals with capital cost reduction by the introduction of low-cost BSs (Relays, or Femto or Pico BSs), and [4], which provides a comprehensive review of a SON as a major maintenance cost-reduction approach.

The main idea of the proposed solution in this chapter is that in order to cope with spatiotemporal changes in either traffic or the cellular environment, a SON engine will dynamically switch to a suitable NTC, based on an adaptive utility function that incorporates major system objectives, e.g., spectral efficiency, fairness, and power consumption, and it can prioritize among these objectives. To overcome the size and complexity of a holistic CSO problem, we propose exploiting a hybrid approach; i.e., a detailed mathematical system model is first constructed, and extensive system-level simulations are performed to generate the whole solution space for all feasible NTCs, consisting of the number of sectors per site "S", the spectrum reuse factor "F", and the number of SCs per site "R". Since possible combinations of "S", "F" and "R" are not large in a practical cellular system and in fact, only the configurations listed in Fig. 3.3 are technically the most feasible ones, the SON engine can easily and effectively search over this confined solution space and adapt the utility to set an optimization target and switch to the most suitable NTC in a time-efficient manner.

3.2.2 System Model and Holistic CSO Problem Formulation

We consider a generic cellular system model as illustrated in Fig. 3.2. We divide the whole area to be optimized by the SON engine into set of Q bins denoted by Q, where q denotes the q^{th} bin, such that $\sum_{q=1}^{Q} a_q = A$, and $\frac{A}{Q} = a_q, \nabla q \in Q$ where A is the total area and area a of the bin is so small

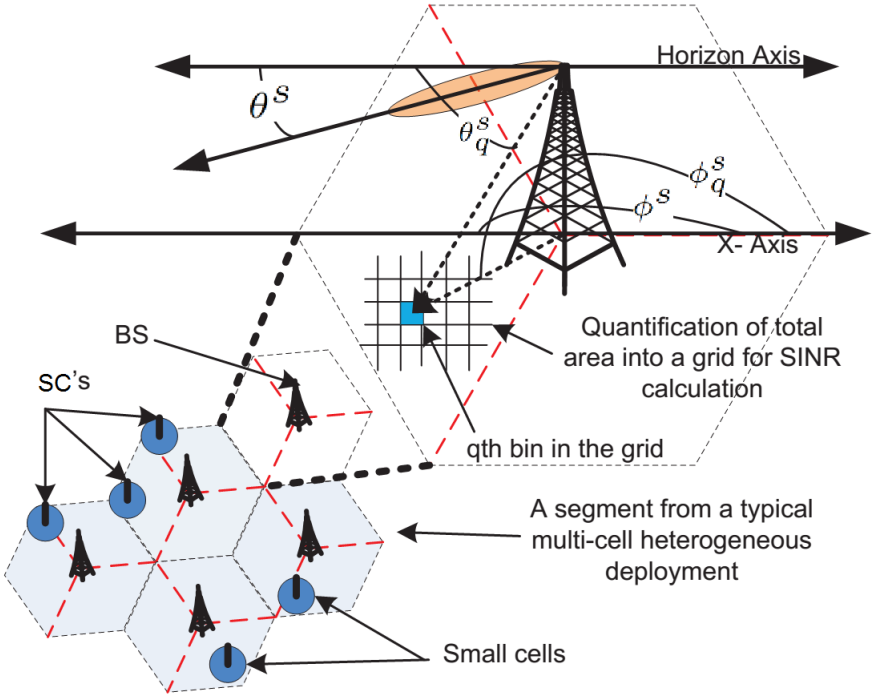


Fig. 3.2: Generic system model used for SINR calculation

that shadowing and path loss can be considered constant within it. Now, using the notation defined in the list of key symbols, the problem of holistic joint optimization of the three performance objectives identified above can be formulated as a multi-objective optimization problem:

$$\max_{Q_b,Q_r,H_r,H_s,S,R,P^s,P^r,\Upsilon_f,\phi,\theta} f(\Upsilon,\Lambda,\Omega) \tag{3.1}$$

subject to feasibility and range constraints on the optimization parameters. The definitions of the parameters in (3.1) are presented in the list of key symbols.

The expression in (3.1) is a holistic CSO problem in which the location of BS and SC, the number of sectors per BS, the number of SCs per BS, the antenna heights, the transmission powers, the antenna azimuth, the antenna tilts, and the frequency reuse have to be optimized to achieve the best possible performance in terms of all three KPIs. Sub problems of such a CSO prob-

lem have been shown to be NP-hard in a number of studies [41, 44, 45, 46]; therefore, metaheuristic techniques are generally utilized to partially explore the solution space of the CSO problem in order to find an acceptable solution. From (3.1), we can obtain some useful insights into the solution space of the problem. Let's take a simple example of only $19 \times 3 = 57$ sectors CS and focus on solving for only one NTC parameter, e.g., the optimal sector azimuth angle. With an over-simplified assumption that the azimuth can only take 10 possible values centered around the nominal azimuth of the sector, a brute force-based solution will have to search among 10^{57} possible azimuth angle combinations. If a system-level evaluation of the KPIs of interest, as a function of azimuth angles, that is generally carried through a simulation tool takes time τ_e (this can be in order of minutes), then finding an optimal solution may take as long as $\frac{10^{57}}{1/\tau_e}$ minutes. However, the actual size of the solution space of a typical holistic CSO problem represented by (3.1) is far larger.

If we apply one of the aforementioned evolutionary metaheuristics used in literature [18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 39, 41, 42, 43], the search space C_p of the holistic CSO problem in (3.1) can be reduced by a factor ϵ. The solution would still require time $\tau = \frac{\left(\frac{C_p}{\epsilon}\right)}{1/\tau_e}$ minutes and yet may not be guaranteed to be optimal. Contrary to most of the works in open literature on CSO, which propose variations and combinations of different metaheuristics to only increase ϵ to reduce the solution time τ, the framework we present in this chapter exploits a bi-pronged approach for increasing the efficiency of the CSO process by reducing both τ_e and C_p instead. First, through the PCF, it eliminates the need for a dynamic simulator for KPI evaluation at each iteration of a search. This is expected to substantially reduce τ_e which will ultimately reduce the τ irrespective of the metaheuristic used to factorize C_p

by ϵ. Second, different parameters have different significance in CSO. Building on further insights into this observation provided by the PCF, we propose a simple algorithm for the holistic CSO problem that can substantially reduce the C_p itself. By allowing conventional metaheuristics to be more thorough, this bi-pronged approach can improve the quality of solutions obtained, while significantly reducing the complexity of the holistic CSO problem.

3.3 Performance Characterization Framework

In this section, we derive quantitative measures for the three KPIs of interest, namely, Υ, Λ, and Ω, in terms of the key NTC parameters, which can be evaluated with low complexity, i.e., without resorting to black-box-type complex dynamic simulators. The SINR perceived in the q^{th} bin from the s^{th} sector (see Fig. 3.2) can be given as follows:

$$\gamma_q^s = \frac{p^s G_q^s \alpha (d_q^s)^{-(\beta)} \delta_q^s}{\sigma^2 + \sum_{\nabla s' \in S} (p^{s'} G_q^{s'} \alpha (d_q^{s'})^{-(\beta)} \delta_q^{s'} . u(\Upsilon_f))} \tag{3.2}$$

where $\{s,\, s'\} \in S, q \in Q$ and $u\,(\Upsilon_f)$ is a unit function that determines whether or not the q^{th} bin will receive interference from a particular sector depending on the frequency reuse. Note that we assume a full load scenario; i.e., all sub-carriers allocated to a cell are simultaneously under use. With this assumption, in calculating SINR, the impact of dynamic scheduling can be omitted, and only static frequency reuse that is part of NTCs can be used to determine the inter-carrier collision and hence interference at a given location. Here, d_q^s is the distance between the q^{th} bin and the s^{th} sector antenna located in the $q_b^{th} \in \mathcal{Q}_b$ bin, given by:

$$d_q^s = \sqrt{(x_{q_b} - x_q)^2 + (y_{q_b} - y_q)^2 + (h_s - z_q)^2} \tag{3.3}$$

Three dimensional (3D) antenna gain can be modeled as in [47] :

$$G_q^s = G(\zeta, D) \times 10^{-1.2(\lambda_v(\frac{\theta_q^s - \theta^s}{\varphi_v})^2 + \lambda_h(\frac{\phi_q^s - \phi^s}{\varphi_h^s})^2)} \tag{3.4}$$

where θ_q^s is the vertical angle in degrees in the s^{th} sector from the reference axis to the q^{th} bin and it can be given as (see Fig. 3.2). The ϕ_q^s is the horizontal angle in degrees in the s^{th} sector to the q^{th} bin with respect to a positive x-axis. λ_h and λ_v represent the weighting factors for the horizontal and vertical beam patterns of the antenna in a 3D antenna model [48], respectively. As indicated in (3.4), the maximum antenna gain G is a function of antenna efficiency ζ and directivity D and it can be written as $G = \zeta D$ where D can be further approximated as: $D = \frac{4\pi}{\varphi_h^s \varphi_v}$.

For the practical cellular antennas, the relationship between the horizontal beam width of the sector antenna and the number of sectors S_b per b^{th} BS site can be modeled as $\varphi_h^s = \frac{\mu * 360}{S_b}$, where μ is a factor representing the overlap between the sectors. Thus using (3.4) in (3.2), the SINR can be determined as in (3.5):

$$\gamma_q^s = \frac{p^s \alpha (d_k^s)^{-(\beta)} \delta_q^s . (\frac{4\pi\zeta}{(\frac{\mu*360}{S_b}\varphi_v)}) . 10^{-1.2(\lambda_v(\frac{\theta_q^s - \theta^s}{\varphi_v})^2 + \lambda_h(\frac{\phi_q^s - \phi^s}{(\frac{\mu*360}{S_b})})^2)}}{\sigma^2 + \sum_{\nabla s' \in S} (p^{s'} \alpha (d_q^{s'})^{-(\beta)}) \delta_q^{s'} . (\frac{4\pi\zeta}{(\frac{\mu*360}{S_b'}\varphi_v)}) . 10^{-1.2(\lambda_v(\frac{\theta_q^{s'} - \theta^{s'}}{\varphi_v})^2 + \lambda_h(\frac{\phi_q^{s'} - \phi^{s'}}{(\frac{\mu*360}{S_b'})})^2)} . u(\Upsilon_f)} \tag{3.5}$$

As desired, the SINR in (3.5) is a function of the key parameters of a given NTC. Similarly, the SINR from the r^{th} SC in the q^{th} bin can be given as follows:

$$\gamma_q^r = \frac{p^r \alpha \left(d_q^r\right)^{-(\beta)} \delta_q^r}{\sigma^2 + \sum_{\forall r' \in \mathcal{R}} \left(p^{r'} \alpha \left(d_q^{r'}\right)^{-(\beta)} \delta_q^{r'}\right)} \tag{3.6}$$

where $\{r, r'\} \in \mathcal{R}$ and $q \in \mathcal{Q}$. Note that for a SC, the antenna gain can be assumed as unity, therefore, it is omitted in the SINR expression. Also, due to

the fact that BSs have much higher Tx powers than SCs, SCs have to duplex with BSs in time or frequency to avoid excessive interference from BSs. With this assumption, only interference from other SCs is considered in (3.6). A frequency reuse of one is assumed among SCs; therefore, no exclusive term to capture the frequency reuse, as in (3.2), is needed in (3.6).

3.3.1 Quantifying Υ–Reflecting Capacity-Wise Performance from a CSO Perspective

We propose a metric, namely, effective spectral efficiency (ESE) to quantify capacity-wise performance denoted by Υ. This metric has semantics similar to the area spectral efficiency, but it does not require a throughput estimation for its calculation; rather, it can be determined through a simple semi-analytical approach. A key advantage of ESE is that it can also serve as the basis for the calculation of the other two KPIs, namely, Λ and Ω. This is useful in modeling the coupling between these contradicting CSO objectives. Below, we explain calculation of ESE.

Since the sub-carrier bandwidth in emerging CSs (e.g., LTE) is fixed, the throughput on a single sub-carrier in a given BS-user link and hence the total throughput of the system depends on the average achievable modulation coding efficiency (MCE) on each link in the system. Over the long term, the MCE depends on the SINR available on that link, whose long-term average value (in a full load scenario, as assumed above) in turn depends mainly on the NTC, as derived above in (3.5) and (3.6).

Let $\mathcal{L} = \{0, 1, 2, 3, ...L\}$ be the set of modulation coding schemes (MCS) available in the standard under consideration. MCE_l denotes the MCE of the l^{th} MCS; $l = 0$ means a MCS with zero spectral efficiency, i.e., no link, representing outage; and L is the MCS with the highest spectral efficiency.

Invoking the bin grid concept, an easily evaluable metric can be given as follows:

$$\Upsilon_{MCE_e} = \sum_{l=0}^{L} \left(MCE_l \times \frac{Q_l}{Q} \right) \tag{3.7}$$

where

$$Q_l = \sum_{\forall q \in \mathcal{Q}} U_l\left(\gamma_q\right) \tag{3.8}$$

Here, γ_q denotes the SINR perceived in the q^{th} bin from the best serving BS sector or SC (whichever is greater), and the unit function $U_l(\gamma_q)$ is defined as follows:

$$\text{For } l \in \mathcal{L} \setminus \{0, L\} : U_l(\gamma_q) = \begin{cases} 1 & T_l < \gamma_q < T_{l+1} \\ 0 & \text{otherwise} \end{cases}$$

$$\text{For } l = L : U_l(\gamma_q) = \begin{cases} 1 & T_l < \gamma_q \\ 0 & \text{otherwise} \end{cases}$$

$$\text{And for } l = 0 : U_l(\gamma_q) = \begin{cases} 1 & \gamma_q < T_0 \\ 0 & \text{otherwise} \end{cases}$$

T_l is the threshold SINR required to use the l^{th} modulation and coding scheme from set $\mathcal{L}$. T_0 is the threshold of the minimum γ, below which a link cannot be maintained with the lowest modulation and coding pair implemented in the standard, and all such points in the coverage area constitute the outage area. Note that

$$\sum_{l=0}^{L} Q_l = Q \tag{3.9}$$

Effectively Q_l is the number of bins in coverage area in which γ_q meets the threshold required to use the l^{th} modulation and coding scheme. A key advantage of quantifying spatial spectral efficiency in this manner is that it has the potential to reflect geographical areas of high importance with weighting factors to pronounce their importance in capacity optimization and reflect them

in the ESE measure proportionally. This provides freedom to tailor this KPI for the CSO process in order to reflect an operator's policy. To set different coverage priorities for different regions, Q in (3.7) that represents number of bins can be replaced with the sum of weights associated with each bin; i.e.,

$$\Upsilon_{MCE_w} = \sum_{l=0}^{L} \left(MCE_l \times \frac{\sum_{l=l'} w_{q_{l'}}}{\sum_{q=1}^{Q} w_q} \right) \tag{3.10}$$

where Υ_{MCE_w} denotes the weighted average MCE, and w_q denotes the weight assigned to the q^{th} bin in proportion to its relative importance in the area of interest. These weights can thus be used to model the QoS requirements of different demographic groups or to differentiate areas with different user densities. $w_{q_{l'}}$ denotes the weight of the q^{th} bin using the l'^{th} MCS, where $l' \in \mathcal{L}$. If not enough data are available to assign precise weights to individual bins, and if operators in general want to ensure that the spatially fair data rates are available throughout the coverage area, harmonic mean can be used instead of the arithmetic mean in (3.7). Unlike the arithmetic mean, the harmonic mean will aggravate the impact of bins with low spectral efficiency and dampen the impact of bins with high spectral efficiency, while representing the overall spectral efficiency of the system. In this case,

$$\Upsilon_{MCE_h} = \frac{Q}{\sum_{q=1}^{Q} (\frac{1}{MCE_q})}, MCE_q > 0 \tag{3.11}$$

where Υ_{MCE_h} denotes the harmonic mean spectral efficiency in the area of interest, and MCE_q denotes the spectral efficiency achievable in the q^{th} bin based on the SINR γ_q perceived in that bin. Note that unlike Υ_{MCE_e} ,Υ_{MCE_h} cannot take into account the outage in the coverage area. While Υ_{MCE} reflects link spectral efficiencies achievable with a particular NTC and can be used as an aspect of capacity, for a holistic quantification of capacity, an important means of cellular capacity, i.e., spectrum reuse, must also be taken

into account.

Against the backdrop of a need for aggressive frequency reuse, we propose reusing the spectrum within a site. By exploiting the fact that aggressive sectorization can provide significant isolation among cells projected from the same BS, the spectrum can be reused within a site among sectors pointing in opposite directions as well as among alternative sectors pointing in different directions, as illustrated in the various NTCs sketched in Fig. 3.3. To quantify the spectrum reuse gain in capacity obtained from such spectrum reuse, we define Υ_f as the "number of times a spectrum is reused within a site". Thus Υ_f can be calculated as follows:

$$\Upsilon_f = \begin{cases} \rho^b \times \frac{S}{F} + \rho^r \times R & \text{if } \ R > 0 \\ \frac{S}{F} & \text{otherwise} \end{cases} \tag{3.12}$$

where ρ^b and ρ^r are factors with which the spectrum is shared between a BS and a SC such that $\rho^b + \rho^r = 1$. F is the number of parts into which the spectrum allocated to the BS (excluding the spectrum allocated to a SC) is divided. Although intra-site spectrum reuse is expected to increase interference and thus decrease Υ_{MCE}, it would be interesting to investigate how a gain in capacity through a higher Υ_f trades against the loss in capacity due to a lower Υ_{MCE}. To incorporate the impact of both of these factors in the cellular capacity, we define the desired capacity-wise KPI named ESE as follows:

$$\Upsilon = \Upsilon_{MCE} \times \Upsilon_f \tag{3.13}$$

Υ_{MCE} can be modeled using (3.7), (3.10), or (3.11) depending on the CSO objectives and service priorities of an operator. On one hand, Υ_{MCE} effectively reflects the capacity gain via spectral efficiency. On the other hand,

Υ_f essentially reflects capacity gain via spectrum reuse efficiency that might come from intra-site frequency reuse (or inter-site frequency reuse, or even fractional frequency reuse, which is not covered in this contribution). Therefore, Υ quantifies the intended capacity-wise KPI from a CSO perspective by incorporating the effect of key NTC factors.

3.3.2 Quantifying Λ–Reflecting SAF from CSO Perspective

From a CSO perspective, the QoS has two aspects: 1) achievable data rates and 2) the spatial fairness of those rates. An explicit metric to quantify only the second aspect is needed from a CSO perspective, as the first aspect is already covered in our definition of Υ. However, for an appropriate measure of fairness, which has to be used in a CSO process as an optimization objective, we must significantly depart from the conventional notion of fairness that is considered when designing very short time-scale adaptive mechanisms, e.g., scheduling or power allocation. For long-term traffic variations, such short-term dynamics can generally be neglected, as they are averaged out. Therefore, it is fairness in space, rather than classic fairness in time, that is more heavily dependent on NTCs and must thus be considered and evaluated during the CSO. More precisely, this spatial fairness of data means the homogeneity of the level of service that can be provided in the coverage area. We build on derivations in the last section and define a metric to reflect the SAF effectively as the inverse of the standard deviation of the spatial distribution of MCE as follows:

$$\Lambda = 1/\sqrt{\frac{1}{Q}\sum_{q=1}^{Q}\left(MCE_q - \sum_{l=0}^{L}\left(MCE_l \times \frac{Q_l}{Q}\right)\right)^2} \tag{3.14}$$

Note that, similar to ESE, SAF can also be evaluated using the SINR expressions derived above. Having an explicit spatial connotation instead of a

temporal one, SAF assigns the cell edge users judiciously higher importance because more bins lie farther from the cell center. Thus, the advantage of SAF is that it is capable of explicitly capturing the cell-center and cell-edge rate disparity. In case, a finite bound-based estimation of SAF is required, Jain's fairness index can also be adapted to estimate the fairness of the service area as follows [49] :

$$JSAF = \frac{\left(\sum_{q=1}^{Q} MCE_q\right)^2}{N \sum_{q=1}^{Q} \left(MCE_q\right)^2} \tag{3.15}$$

3.3.3 Quantifying Ω–Reflecting Power Consumption Wise Performance from a CSO Perspective

Power consumption in a cellular system has many complicated and interrelated components. Considering the scope of this contribution, we focus on the five selected elements of NTCs that mainly determine the power consumption of a given NTC, i.e., types of access points (BS or SC), number of sectors per site, number of SCs per site, transmission powers and sector overlap. These are the main parameters that make the power consumption in various cellular systems' NTCs different from each other. To that end, we model the power consumption on a site while incorporating both the fixed and variable power consumption per site that in turn depends on the type of BS. Fixed power consumption is the power that is consumed in keeping the circuitry of BS sectors alive regardless of whether there is traffic or not. Fixed power remains non-zero until all sectors and SCs associated with a BS are completely switched off. Variable power consumption is the power required for transmission on air interface, and it varies with the traffic load. Total power consumption in the b^{th} BS site (including that of all sectors and associated SCs) can thus be written as (3.16) below:

$$p = \sum_{s=1}^{S_b} \left\{ p_f^s + p_v^s \left(G\left(\zeta^s, D^s\right), p_t^s, \eta^s \right) \right\} + \sum_{r=1}^{R_b} \left\{ p_f^r + p_v^r \left(G\left(\zeta^r, D^r\right), p_t^r, \eta^r \right) \right\} \tag{3.16}$$

where subscripts f, v, and t denote fixed, variable, and transmission powers respectively. For the sake of simplicity, we do not consider any stray losses, e.g., feeder loss and connector loss, as they are negligible for the purpose of this analysis. The variable power consumption within each sector or SC further depends on the transmission power p_t^s and p_t^r, the traffic loading factors for sectors and SCs (between 0 to 1) η^s and η^r, respectively, and antenna gain G of the sectors and SCs respectively. Furthermore, antenna gain is a function of antenna efficiency ζ and directivity D. The directivity of the antenna determines its gain and hence the transmission power required to provide a certain coverage and service level. For almost all commercial antennas used in cellular systems, the directivity can be approximated as follows [50]:

$$D \approx \frac{4\pi}{\varphi_h \varphi_v} \tag{3.17}$$

In commercial cellular systems, the typical vertical beam width of an antenna is approximately $\varphi_v \approx \pi/18$ radians, and the horizontal beam width depends on the number of sectors per access point. For a BS with three sectors and six sectors, beam widths of around 70^o and 35^o respectively are generally used. Using μ, defined above as the factor determining the overlap between the adjacent sectors, we can write the horizontal beam width as a function of S_b as $\varphi_h = 2\pi\mu/S_b$. Using these values of φ_h and φ_v, (3.17) can be written as follows:

$$D \approx \frac{36 S_b}{\mu\pi} \tag{3.18}$$

The typical value of μ can be assumed to be $\mu = 1.1$. To achieve a desired

effective isotropic radiated power (EIRP) in the coverage area, less transmission power p_t will be required for antennas with higher gains, as indicated below:

$$EIRP = \zeta \times D \times p_t \tag{3.19}$$

If p_d is the power required to achieve the desired $EIRP_d$ with an omnidirectional antenna, then

$$p_d = \frac{EIRP_d}{\zeta D} \tag{3.20}$$

Therefore, for a given coverage level, if more sectors per site are used, then less transmission power per sector would be required due to high directivity and hence higher gains of the antennas. The variable circuit power per sector for the desired $EIRP_d$ can thus be written in dB as follows:

$$p_v^s = 10\log_{10} p_d^s - 10\log_{10}\left(\frac{2\zeta^s S_b}{\mu\varphi_v}\right) + 10\log_{10}\eta^s \tag{3.21}$$

Similarly, the variable circuit power on an SC can be written as follows:

$$p_v^r = 10\log_{10} p_d^r - 10\log_{10}\left(\frac{2\zeta^r}{\varphi_v}\right) + 10\log_{10}\eta^r \tag{3.22}$$

Substituting (3.21)-(3.22) to (3.16) and re-arranging, we obtain (3.23):

$$\Omega = \left(\sum_{s=1}^{S_b}\left\{p_f^s + \mu\left(\frac{\eta^s\varphi_v^s p_d^s}{2\zeta^s S_b}\right)\right\} + \sum_{r=1}^{R}\left\{p_f^r + \frac{\eta^r\varphi_v^r p_d^r}{2\zeta^r}\right\}\right) \tag{3.23}$$

$$\frac{\Omega}{\Upsilon} = \frac{\left(\sum_{s=1}^{S_b}\left\{p_f^s + \mu\left(\frac{\eta^s\varphi_v^s p_d^s}{2\zeta^s S_b}\right)\right\} + \sum_{r=1}^{R}\left\{p_f^r + \frac{\eta^r\varphi_v^r p_d^r}{2\zeta^r}\right\}\right)}{\sum_{l=0}^{L}\left(MCE_l \times \frac{Q_l}{Q}\right) \times \Upsilon_f} \tag{3.24}$$

Equation (3.23), on one hand, provides a simple metric to quantify the power consumption in an NTC as a function of the number of sectors per site, the number of SCs per site, transmission powers, sector overlap, and antenna

Table 3.1: Modeling parameters

Parameters	Values
System topology	19 sites (1-6 sector/site)
BS transmission power	39 dBm
BS Inter site distance	1200 meters
BS height	32 meters
User antenna	0 dB (Omini directional)
BS antenna vertical beam width	10^0
BS antenna horizontal gain weight	0.5
BS antenna vertical gain weight	0.5
BS antenna maximum gain	18 dB
BS antenna maximum attenuation	20 dB
Frequency	2 GHz
Path loss model	Cost Hata
Shadowing STD	8 dB

beam widths. This metric also takes into account an additional factor, namely, the traffic load factor, which is not a direct part of NTCs but can heavily affect the power consumption. The split ratio between the fixed power consumption and transmission power can be used to model various BS types as well. On the other hand, equation (3.24) provides a metric to quantify the long-term average energy consumption in $\frac{Joules}{bits}$ for a given NTC.

3.4 Performance Evaluation of Different NTCs

In this section we evaluate the performance of a range of potential NTCs using the PCF.

3.4.1 System Model for Performance Evaluation

A total of 26 NTCs with generally feasible combinations of key NTC parameters F, S, and R (see Fig. 3.3) are evaluated while other parameters are kept fixed at the values listed in Table 3.1. Two tiers of cells are modeled for each NTC to consider a realistic amount of interference in a multi-cellular scenario. Shadowing and appropriate path loss models for BSs and SCs similar to [51] are used to model a realistic cellular system environment. In SC-CS, SCs are located at half of the inter-site distance, where the SINR is minimum, i.e., where the far end corners of adjacent sectors join. To map the SINR γ_q to the long-term average link spectral efficiency, we refer to SINR thresholds for

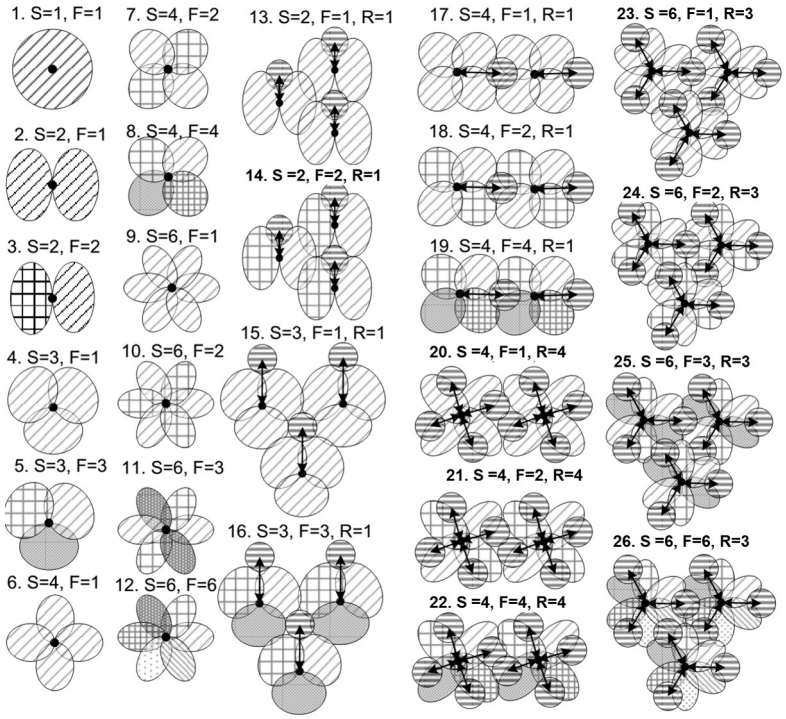


Fig. 3.3: Twenty-six different NTCs with varying S, F and R, which are investigated in this chapter. Dots in the center of each site represent base station locations. Oval shapes represent sectors, and small circular shapes represent small cells attached to a site. Filling patterns represent the frequency reuse pattern whereas arrows represent backhaul links between base stations and small cells.

MCSs used in LTE. A given NTC is denoted by the number of sectors per site S, frequency reuse F, and the number of SCs per site R. Thus, e.g., an NTC denoted by "25. $S = 6$, $F = 3$, $R = 3$" means that NTC no. 25 has six sectors per site, and the spectrum allocated to BS (after splitting with SC) is divided into three equal parts; each part is allocated to three adjacent sectors, and the pattern is repeated for the other three sectors on the site such that sectors using the same spectrum are pointing in opposite directions to each other and the site has three SCs. Thus, in NTC 25 the spectrum is reused $\Upsilon_f = \rho^b \times \frac{S}{\mathrm{F}} + \rho^r \times R = 0.5 \times \frac{6}{3} + 0.5 \times 3 = 2.5$ times within a site area. For brevity, the analysis hereafter will use Υ_{MCE_e} as a measure of capacity and Λ as a measure to reflect SAF.

3.4.2 Analyzing Capacity Wise Performance

Fig. 3.4 plots Υ, Λ, and Ω evaluated for all 26 NTCs under consideration, normalized by their maximum values. It can be seen that different NTCs offer different trade-offs among different KPIs. For ease of discussion while probing into these trade-offs, we first focus on NTCs 9-12, all with $S = 6$. It can be seen that from NTC = 9 to NTC = 12, as frequency reuse is made less tight with other parameters being fixed, the overall capacity of the system, i.e., Υ, still increases (see Fig. 3.4). This is because the increase in Υ_{MCE_e} due to decreased interference overweighs the loss in Υ_f. As a net result, Υ is hence larger in NTC = 10, 11, 12 compared to NTC = 9. However, there is a payoff of this gain. It can be seen that Λ (i.e., SAF) continuously decreases from NTC = 9 to NTC = 12. The reason for this will be discussed in the next subsection. By comparing the Υ for SC-CS with that for CS, it can be easily seen that SCs bring a significant improvement in overall capacity. There are two reasons for this improvement. First, the much smaller height and lower

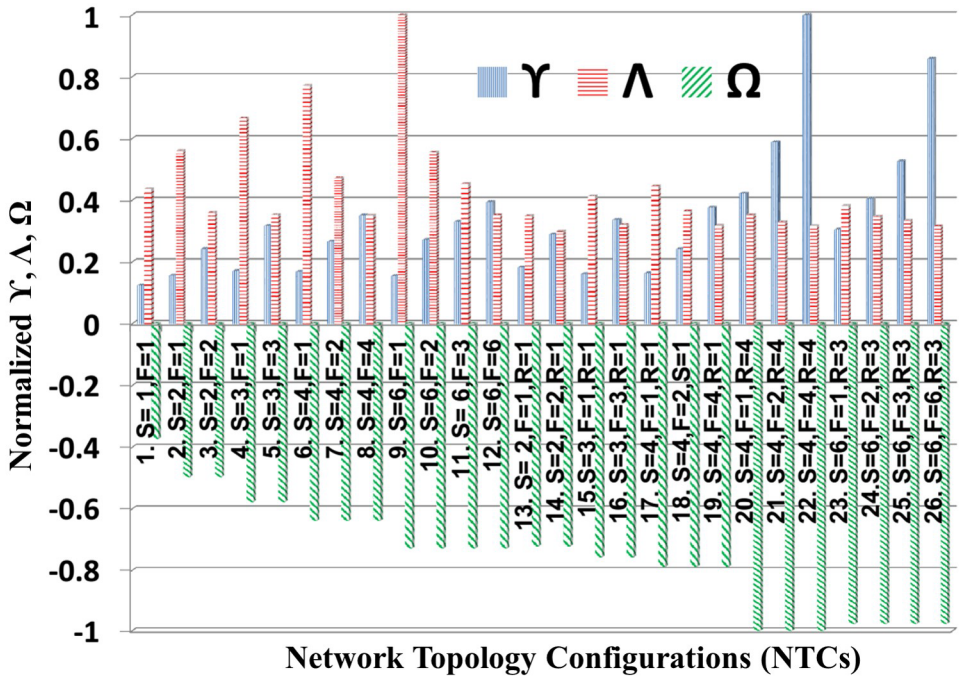


Fig. 3.4: Comparison of different NTCs in terms of their capacity Υ, service area fairness Λ and power consumption Ω

transmission power of SC causes and suffer from much lesser interference, resulting in a better Υ_{MCE_e}. Second, in addition to a higher Υ_{MCEe}, there is another positive contribution of SCs towards a higher Υ that is explained as follows: Let us assume that three SCs are working in a cell. In this case, the spectrum is divided into two parts for sharing between a BS and SC. This reduces Υ_f by half only, compared to a scenario with three sectors where Υ_f will be reduced by a factor of 3. These two reasons together make an SC a more advantageous method to boost capacity, compared to adding more sectors. However, there is a payoff for this gain in capacity achieved by SC in terms of both SAF as well as power consumption. It can be seen from Fig. 3.4 that SC-CS in general has a lower SAF and higher power consumption compared to CS.

3.4.3 Analyzing SAF

From the results in Fig. 3.4, it can be noted that the SAF increases with an increase in the number of sectors, but it decreases with an increase in F (or in other words, a decrease in Υ_f). This is because increasing the number of sectors in general decreases the cell edge interference, thereby making the geographical distribution of data rates more uniform in a cell. A low Υ_f means less intra-site reuse, and interference consequently comes primarily from adjacent sites rather than adjacent sectors, leading to classic scenarios where cell edge experience much more interference and hence a lower SAF than cell center users. On the other hand, SAF in SC-CS is noticeably lower than that in CS due to the drastic change in distribution of data rates brought by SCs.

3.4.4 Analyzing Power Consumption

It is clear from results in Fig. 3.4 that, as expected, total power consumption increases as both the mean of increasing capacity, i.e., sectors, or SCs are added. Therefore, even though SCs offer suitable means to increase capacity, as seen above, a slightly higher power consumption is another payoff for them in addition to a poorer SAF. Fig. 3.5 plot the total power consumptions for a range of R and S using (3.23) and the preceding analysis. $\eta^s = \eta^r = 1$ is assumed because we are considering a full load scenario. The antenna efficiency of commercial antennas is used, i.e., $\zeta = 60\%$. $P_f^s = 15W$, with $P_f^r = 0.5\,P_f^s$ is used for the reasons explained in [14]. It can be seen that in addition to the spectral efficiency and spatial fairness of data rates, power consumption also varies with S and R, and it thus adds a third dimension to the capacity-QoS trade-off in dimensioning NTCs. Fig 3.5 illustrates that power consumption per site increases more rapidly with an increase in the

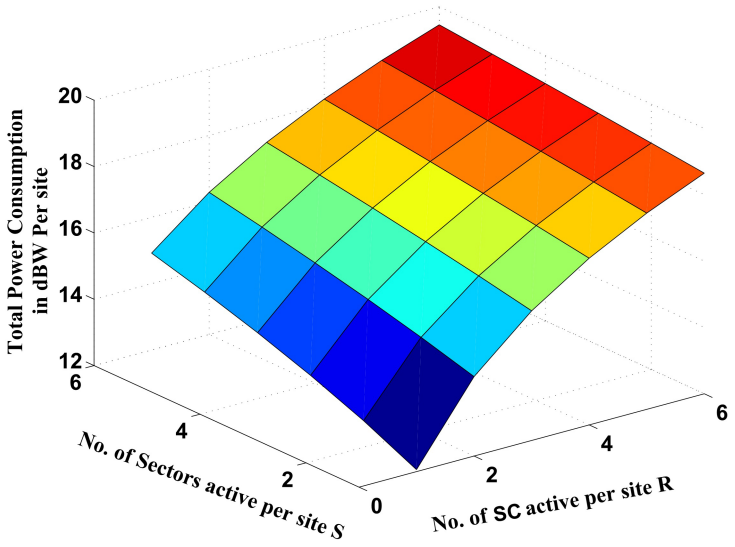


Fig. 3.5: Total power consumption per site

number of SCs (i.e., R) than an increase in the number of sectors per site (i.e., S). This is mainly because each SC has an omnidirectional antenna, so there is no compensating factor as in the case of sectors.

3.4.5 Trade-Off between the Three Performance Aspects

Finally, from the results in Fig. 3.4, it can be seen that no single NTC is simultaneously optimal in all three performance aspects. Here, the key observation is that there exists a certain pareto optimality in which one objective generally improves only with a loss in another. Therefore, the PCF's capability to precisely quantify this trade-off with computation efficiency can actually help to design an NTC that is optimal for simultaneously meeting the multiple CSO objectives in an operator's intended order of priority. Although a regular topology has been assumed in the analysis for the sake of simplicity, in cases of realistic irregular topologies, the PCF can build on the real user MDT reports, and the KPIs produced by the PCF and hence the optimal NTCs determined will consequently still be optimal for real networks.

3.5 Application of PCF in Holistic Optimization

Despite the fact that the PCF can reduce τ_e and thus can reduce the overall solution time, the holistic optimization of all parameters together remains a daunting task. In this section, we propose a simple three-step heuristic algorithm for a SON engine to simplify the holistic optimization by using the PCF.

3.5.1 Classify Parameters, Prioritize Objectives and Solve Subproblems: A Pragmatic Heuristic for Holistic Optimization

The CPS algorithm has the following three steps:

1. Classify the parameters of interests into hierarchical groups based on their impact on the KPIs. For example, parameters that substantially determine network performance can be classified into a group named gross parameters (GPs), and parameters that fine-tune network performance can be placed in another group called fine tuning parameters (FTPs). This grouping can be done by examining the role of a particular parameter through the PCF.
2. Prioritize the objectives, which involves modeling the optimization objective of the holistic CSO problem using the PCF. This modeling should reflect the operator's priorities for each KPI. This step will be explained in detail through a case study below.
3. Solve the subproblem.
   (a) Starting with the highest group in the parameter hierarchy (resulted from step 1), optimize the objective function defined in step 2 for the parameters in this group, considering it as a subproblem that is independent of the groups below it.

i. To solve this optimization sub-problem, normalize the KPIs to make them unitless in order to bring them to the same scale.

ii. Use these normalized values of the KPIs in the objective function defined in step 2 and solve the subproblem using an exhaustive search or metaheuristics depending on the parameter group size.

(b) Once a group of parameters is optimized, lock all parameters in that group at their optimal values, and repeat step 3 for the lower groups until all groups are optimized.

The CPS algorithm is further explained below through a case study.

3.5.2 A Case Study for CPS

As a case study, we consider the joint optimization of four key NTC parameters, namely F, S, R, and θ, which have been largely overlooked in the literature. The CSO problem under consideration can thus be written as follows:

$$\max_{F,S,R,\theta} \left\{ \Upsilon\left(F,S,R,\theta\right), \Lambda\left(F,S,R,\theta\right), \Omega\left(F,S,R\right) \right\} \tag{3.25}$$

From the previous section, we know that no single NTC is optimal for Υ, Λ, and Ω simultaneously. This also implies that (3.25) is non-convex and hence difficult to solve with analytical approaches. Below, we apply the CPS to find a solution with low complexity. We place F, S, and R in the GPs group and θ in the FTP group. This grouping is quite intuitive and can also be inferred from the expressions in (3.5) and (3.6), which demonstrate that F, S and R have a more profound impact on the SINR and hence the KPIs associated with it than θ.

Optimizing GPs

The GP optimization problem can be written as follows:

$$\max_{F,S,R} \{\Upsilon(F,S,R), \Lambda(F,S,R), \Omega(F,S,R)\} \tag{3.26}$$

Since the mutual priority of these objectives and their target values are strongly dependent on the operator's policy [4], we propose using multi-objective optimization, as used in [52] by representing the three objectives simultaneously as a single utility function; i.e.,

$$v = \begin{cases} v_g(\Upsilon, \Lambda, \Omega), & \text{General Optimization} \\ v_t(\Upsilon, \Lambda, \Omega), & \text{Targeted Optimization} \end{cases} \tag{3.27}$$

where the subscripts g and t denote the general and targeted cases respectively as further explained below:

1. Case 1 (General Optimization): This case represents a scenario in which the operator has no specific target values for the KPIs but has a certain priority for each KPI. In this case, the optimization problem can be modeled as:
$$\max_{F,S,R} v_g(\Upsilon, \Lambda, \Omega) = \max_{F,S,R} (\lambda_1 \Upsilon + \lambda_2 \Lambda - \lambda_3 \Omega) \tag{3.28}$$
This utility function can reflect the mutual priority among these objectives. Below, we present some exemplary rules to manifest these priorities:
   (a) If the operator has equal priority for all the KPIs, then, in (3.28), set the following
   $$\lambda_1 = \lambda_2 = \lambda_3 = 1/3 \tag{3.29}$$
   (b) If the operator wants to maximize a specific objective (the d^{th} objective), while neglecting others, then in (3.28), set the following

$$\lambda_i = \begin{cases} 1 & \text{if } i = d \text{ , i = 1, 2, 3} \\ 0 & \text{otherwise} \end{cases} \tag{3.30}$$

(c) If the operator has a specific priority for each objective, it can be represented by weights such that

$$\lambda_1 + \lambda_2 + \lambda_3 = 1 \tag{3.31}$$

2. Case 2 (Targeted Optimization): This case represents the scenario where the operator has specific target values to be achieved in each performance aspect. In this case, the optimization problem can be written as (3.32):

$$\min_{F,S,R} \upsilon_t \left(\Upsilon, \Lambda, \Omega\right) = \min_{F,S,R} \left| \sqrt{\lambda_1 \left(\Upsilon - \Upsilon_t\right)^2 + \lambda_2 \left(\Lambda - \Lambda_t\right)^2 + \lambda_3 \left(\Omega - \Omega_t\right)^2} \right| \tag{3.32}$$

The rules for utility adaptation are as follows:

(a) If the operator wants to achieve desired targets in each metric with the same priority, then substitute (3.29) in (3.32).

(b) If the operator has a desired target value in one objective but has no priority in others, then substitute (3.30) in (3.32).

(c) If the operator has specific values of each metric as targets but a different priority for each target to be met, then substitute (3.31) in (3.32).

Fig. 3.4 provides the solution space for the problem in (3.26), obtained by the normalization of the KPIs with their respective maximum values.

Fig. 3.6 plots utility υ_g for four sets of different objective priorities. With an equal priority of all three objectives, we can see that the GP values in NTC = 9 are optimal. When capacity has the highest priority i.e., 80%, and fairness and power consumption have lower and equal priorities of 10% each,

the GP values in NTC = 22 are optimal. On the other hand, when fairness has highest importance, i.e., 80%, and capacity and power consumption have lower and equal priorities of 10% each, the GP values in NTC = 9 become optimal. When power consumption is the most important target, with an 80% importance factor, and fairness and spectral efficiency are lower priorities, with an importance of just 10% each, the optimal GP choice is given by NTC=1.

Fig. 3.7 plots υ_t for three different sets of target values of the three objectives, each having the same priority, i.e., $\lambda_1 = \lambda_2 = \lambda_3 = 1/3$. The first case (blue) represents the CSO scenario when the operator wants both capacity and fairness-wise performance to be closest to their absolute optimal values but has some flexibility in power consumption. In this case out of the 26 GP combinations explored, the optimal solution is NTC = 4. The second case (red), represents a scenario in which the power is needed to be closest to optimal, followed by spectral efficiency, and finally fairness. Now the NTC = 5 can be seen to be the optimal solution. The last case (green) represents scenarios where the operator wants SAF to be closest to its absolute optimal and can tolerate middle level performance in capacity, followed by power consumption. In this case, NTC = 9 provides the optimal GP values to meet these priorities.

Optimizing FTPs:

Assuming the operator's business model requires all three KPIs to be equally important, this policy will be modeled with utility 1, with $\lambda_1 = \lambda_2 = \lambda_3 = 1/3$. In this case, the GP optimization, i.e., the solution to the subproblem in (3.26), will return a solution ($F = 1$, $S = 6$, $R = 0$). The next step in the holistic CSO problem, according to the CPS algorithm, can now be written as follows:

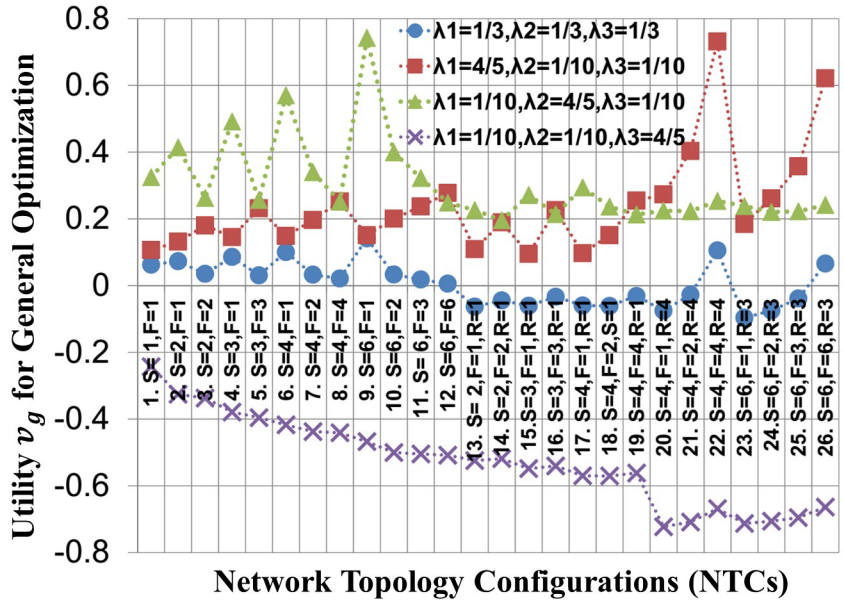


Fig. 3.6: Solution space for general optimization

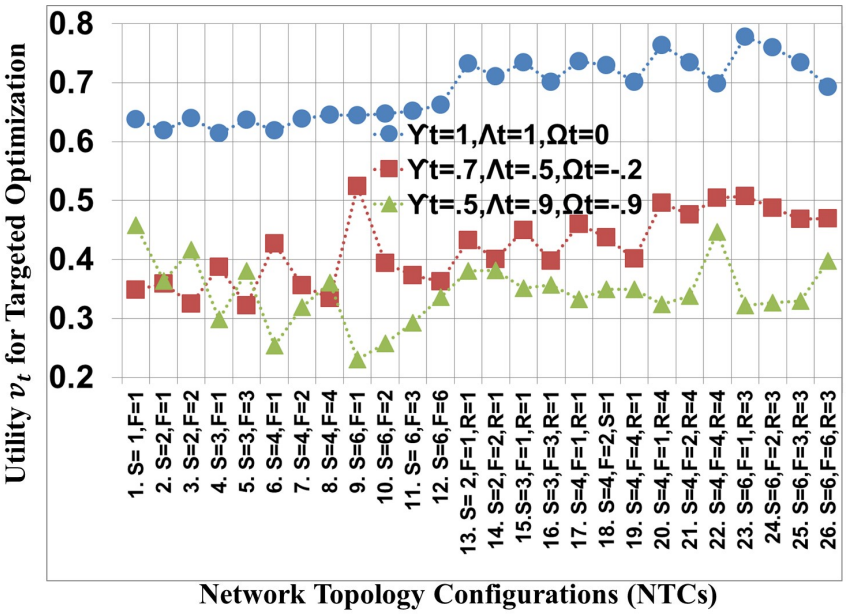


Fig. 3.7: Solution space for targeted optimization

$$\max_{\theta} \{\lambda_1 \Upsilon(\gamma(\theta)) + \lambda_2 \Lambda(\gamma(\theta))\} \tag{3.33}$$

Note that since Ω is not a function of θ, it does not have to be included in the optimization problem. The KPIs Υ and Λ are functions of SINR γ, which is a further function of θ, which is a vector of the tilt angles of all sectors in the system, as modeled in (3.5). Note that in our particular case, each site has the same F and S. Therefore, from the insights obtained from (3.5), it is clear that optimal tilt angles, being dependent on height as well as S and F (GPs), will be the same across the network. With this additional simplification, the PCF can be used to quickly draw the solution space of (3.33), which is presented in Fig. 3.8. It can be seen that, again, there is a strong trade-off between the two KPIs, and no single tilt is optimal for both Υ and Λ. Using the same utility-based approach as proposed above, namely the optimal value of the FTP that meets the operator's defined objective, the solution can be easily found. For example, in a case where $\lambda_1 = \lambda_2$ reflects the operator's priorities, the solution is $\theta = 14^0$. The solution to our CSO problem in (3.25), for given KPI priorities set by the operator, is thus $(F = 1, S = 6, R = 0, \theta = 14)$.

3.5.3 Complexity of PCF and CPS based holistic CSO approach

Since the grouping of parameters substantially reduces the search space size, and the PCF reduces τ_e compared to traditional dynamic simulation-based SON approaches, the CPS algorithm can greatly reduce the solution complexity of the holistic CSO problem. More specifically, if a conventional approach takes time $\tau = \frac{\left(\frac{V^M}{\epsilon}\right)}{1/\tau_e}$ to solve the CPS problem with M optimization parameters, each of which can take V different values, then the CPS will take the following amount of time:

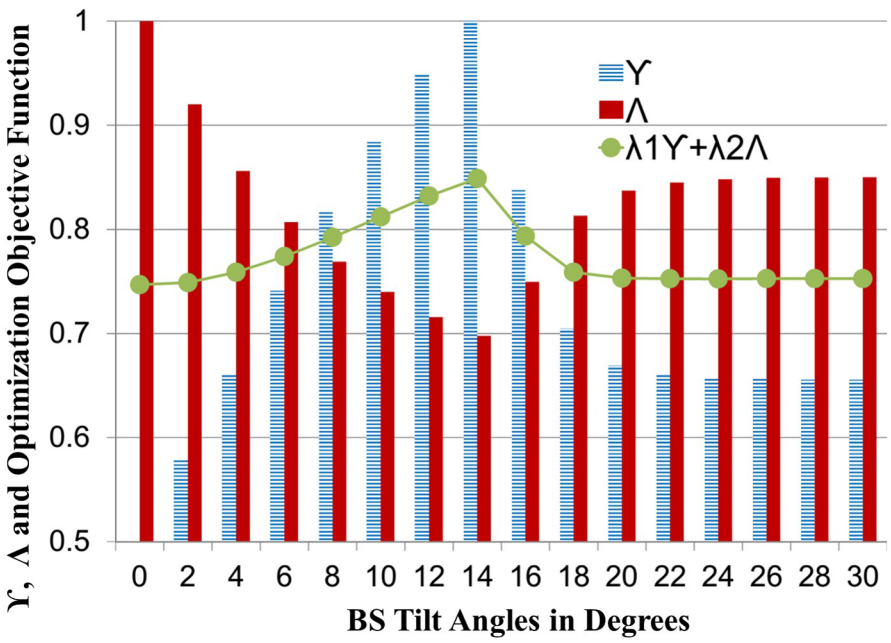


Fig. 3.8: Υ, Λ and optimization objective function of tilt angle, for NTC = 9

$$\tau' = \frac{\tau}{\left(\frac{V^M}{\sum_{i=0}^{G}(V^{\frac{M}{g_i}})} \times \frac{\tau_e}{\tau'_e}\right)} \tag{3.34}$$

where τ'_e is the time required for the individual evaluation of KPIs using the PCF and G is the number of groups in which CPS divides the parameters. This generally implies that $\tau >> \tau'$. For our particular case study, the feasible combinations of F, S, and R were as low as 26, and τ'_e on a regular desktop computer was less than 1 second. Therefore, it took less than a minute to explore the search space for GPs and almost the same time for FTPs.

3.6 Conclusion

This chapter presented a framework to quantify, analyze, and optimize the three major KPIs–capacity, SAF and power consumption–used for the holistic optimization of SON-enabled heterogeneous cellular systems. The PCF proposed in this contribution can model the KPIs of interest as functions of

a comprehensive set of optimization parameters such as the spectrum reuse factor, the number of sectors per site, the number of SCs per site, adaptive coding, and modulation. The metrics derived in the PCF can be quickly evaluated semi-analytically, and they can thus facilitate a solution to the multi-objective, holistic optimization problem that is otherwise tackled using black-box type complex dynamic simulation models. Using the PCF, we also evaluated and compared 26 different network topologies and quantified their relative gains. We analyzed the respective trade-offs offered by each NTC in terms of capacity, SAF and power consumption. Our results demonstrated that contrary to common notion, NTCs with the highest spectrum efficiency are not necessarily those that resort to full frequency reuse. The insights obtained by the proposed framework can help to address new requirements from future heterogeneous cellular networks. Building on these insights, we proposed a heuristic CPS algorithm for holistic optimization. Through a case study, we demonstrated how the PCF and CPS together can be used for a wide range of cellular optimization scenarios with low complexity.

CHAPTER 4

Spatiotemporal Mobility Prediction

> The more unpredictable the world is the more we rely on predictions.
>
> Steve Rivkin

In this chapter, we present a contribution in the area of mobility prediction as an enabler of proactive SONs in cellular networks. We develop and analyze a semi-Markov model-based spatio-temporal mobility prediction model. The proposed mobility prediction model overcomes the limitation of conventional DTMC-based prediction models that fail to incorporate the time dimension, i.e., "Time of next HO." Next, we propose a novel method to map the next cell spatiotemporal HO information to the estimated future location coordinates based on the idea of Landmarks. This novel method further increases the spatial resolution of the future location estimation without requiring an increase in the number of states for the semi-Markov model. The accuracy of the proposed model is quantified through experimental evaluation, leveraging real network traces generated by smartphone applications as well as through simulations.

4.1 Introduction

Mobility is the raison d'être of wireless cellular networks. However, the planned design of future wireless networks namely 5G resorts to extreme cell densification. This design is antagonistic to capability of cellular networks for seamlessly supporting user mobility. This challenge has been recognized as

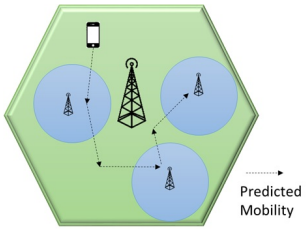


Fig. 4.1: Mobility prediction in cellular network

one of the major hurdles in the successful realization and deployment of 5G cellular technology. This chapter presents a truly revolutionary approach that transforms mobility from a bane of the cellular industry into a blessing. The main idea behind the proposed approach is to first develop robust models to predict certain attributes of user mobility and then exploit these attributes to ultimately develop the foundations for the much needed next generation mobility management proactive SON. This proactivity is achievable through anticipating user behavior and predicting a future network state by exploiting historical network information referred to as big data. Endowed with these proactive predictive capabilities, network resources can be pre-allocated more intelligently and in a more efficient manner than ever before [5].

User mobility prediction is one of the core ingredients of the proactive SON paradigm; it predicts the future locations of users in terms of the associated BSs (see Fig. 4.1). This enables the reservation of network resources in future identified cells for a seamless handover experience as well as for traffic forecasting purposes that drive SON functions like ES, LB etc.

Our rationale for building and utilizing mobility prediction is backed by a landmark study that analyzed real data for 10 million mobile users [53] and revealed that typical human mobility features 93% average predictability. The mobility prediction model developed in this contribution exploits the following idea: transition probability to a next cell can be predicted by modeling user transition from one cell to another as a Markov stochastic process and using

HO history to estimate state transition probabilities. The DTMC has been commonly used in the literature for mobility prediction purposes [54, 55, 56]. Compared to more complex and more space-consuming compression-based predictors, the Markov-based scheme can yield a more scalable solution, as it does not need to store users' past movements. Instead, the crux of this information is captured by transition probabilities. However, the DTMC is memory-less and assumes that sojourn time is geometrically distributed and that each transition takes place in one unit of time. Considering these limitations of the DTMC model, the aforementioned works have utilized the DTMC for only spatial prediction, i.e., the identification of a future cell only, without any information about the time at which a handover may take place. The CTMC is the DTMC's continuous counter part and it can be utilized for mobility prediction if human mobility is assumed to be memory-less and if cell sojourn time is assumed to be exponentially distributed. As per [57], human mobility exhibits a memory property and can be best approximated with power law (heavy tailed) distribution instead of memory-less exponential distributions. The semi-Markov model is fortunately an advanced class of Markov models that allows for arbitrary distributed sojourn times. Few recent works have characterized the prediction accuracy performance of the semi-Markov based model for mobility prediction [58, 59]. However, the aforementioned relevant studies utilized historic, publicly available WLAN traces, not cellular network mobility traces. Those WLAN mobility traces exhibit large sojourn times due to relatively fewer mobility dynamics in WLAN compared to a cellular network. Therefore, the conclusions drawn from these WLAN studies cannot be directly applied to current cellular networks, such as LTE, where the device form factor and user behavior are drastically different than those of WLAN users in 2004. The study presented in this chapter

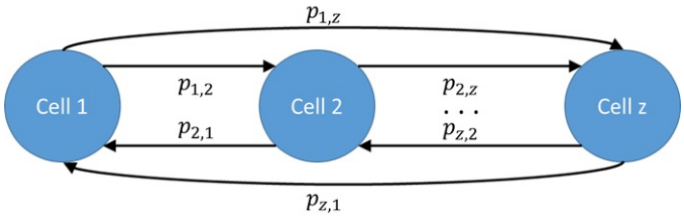


Fig. 4.2: Probability state transition diagram

fills this gap. We gather real LTE network's mobility traces in live cellular network using a state-of-the-art application. These traces exhibit highly dynamic characteristics that are intrinsic to the cellular network, and they thus enable a realistic evaluation of the prediction accuracy of the semi-Markov-based method for cellular network mobility. Another contribution of this chapter is that instead of relying on historic public datasets, we use a novel methodology of employing smartphone applications, based on the idea of participatory sensing, to collect real LTE network data for building, training, and evaluating the performance of mobility prediction schemes in a live network. We also present a method of quantifying the gains of mobility prediction techniques from the perspective of proactive SON-enabled cellular networks.

4.2 Mobility Prediction Model

We begin by modeling user mobility as a semi-Markov renewal process $\{(X_n, T_n) : n \geq 0\}$ with discrete state space $\mathbb{C} = 1, 2, 3 \ldots, z$, where T_n is the time of the n^{th} transition, X_n is the state at the n^{th} transition and the total of z cells [60]. Each cell is represented by the state of the semi-Markov process, and a handover from one cell to another is considered to be a state transition. It is assumed that the process is time-homogeneous during the time period in which the model is built. Fig. 4.2 displays a state transition diagram for the semi-Markov model, wherein $p_{i,j}$ is the probability of transition from cell i to

j. The associated time-homogeneous semi-Markov kernel for user "u" which is the probability of transition to the j^{th} cell if the user has already spent time t in the i^{th} cell, is defined as follows:

$$\psi_{i,j}^{(u)}(t) \;=\; Pr(X_{n+1}^{(u)} = j, T_{n+1}^{(u)} - T_n^{(u)} \leq t | X_n^{(u)} = i) = p_{i,j}^{(u)} \Gamma_{i,j}^{(u)}(t) \tag{4.1}$$

where

$$p_{i,j}^{(u)} \;=\; \lim_{t \to \infty} \psi_{i,j}^{(u)}(t) = Pr(X_{n+1}^{(u)} = j | X_n^{(u)} = i),\; p_{i,j}^{(u)} \in P^{(u)} \tag{4.2}$$

and

$$\Gamma_{i,j}^{(u)}(t) = Pr(T_{n+1}^{(u)} - T_n^{(u)} \leq t | X_{n+1}^{(u)} = j, X_n^{(u)} = i) \tag{4.3}$$

Here, $p_{i,j}^{(u)}$ is the probability of handover of user "u" from cell i to j, $\mathbf{P}^{(u)}$ is the probability transition matrix of the embedded Markov chain of user "u" given as

$$\mathbf{P}^{(u)} = \begin{bmatrix} p_{1,1}^{(u)} & p_{1,2}^{(u)} & \cdots & p_{1,z}^{(u)} \\ p_{2,1}^{(u)} & p_{2,2}^{(u)} & \cdots & p_{2,z}^{(u)} \\ \vdots & \vdots & \vdots & \vdots \\ p_{z,1}^{(u)} & p_{z,2}^{(u)} & \cdots & p_{z,z}^{(u)} \end{bmatrix} \tag{4.4}$$

and $\Gamma_{i,j}^{(u)}(t)$ is the sojourn time distribution of user "u" in cell i when the next cell is j. It is important to note here that the handover from a cell to itself is not allowed; therefore, the diagonals of the matrix $P^{(u)}$ will all be zeros, and the matrix will be a hollow matrix. Furthermore, direct handovers are possible between neighboring cells only. The probability of user "u" in cell i leaving cell i before or at time t, regardless of the next cell is defined as follows:

$$\Lambda_i^{(u)}(t) \;=\; Pr(T_{n+1}^{(u)} - T_n^{(u)} \leq t | X_n^{(u)} = i) = \sum_{j=1}^{z} \psi_{i,j}^{(u)}(t) \tag{4.5}$$

Now the time-homogeneous semi-Markov process of user "u" is defined as $X = (X_t,\, t \in \mathbf{R}_{\mathbf{0}}^{+})$ with state transients as follows:

$$\chi_{i,j}^{(u)}(t) = Pr(X_t^{(u)} = j | X_0^{(u)} = i) \tag{4.6}$$

$$= (1 - \Lambda_i^{(u)}(t))\delta_{i,j} + \sum_{m=1}^{z} \int_0^t \chi_{m,j}^{(u)}(t-\tau) d\psi_{i,m}^{(u)}(\tau) \tag{4.7}$$

$$= (1 - \Lambda_i^{(u)}(t))\delta_{i,j} + \sum_{m=1}^{z} \int_0^t \frac{d\psi_{i,m}^{(u)}(\tau)}{d\tau} \chi_{m,j}^{(u)}(t-\tau) d\tau \tag{4.8}$$

where $\delta_{i,j}$ is the Kronecker function defined as:

$$\delta_{i,j} = \begin{cases} 0 & , i \neq j \\ 1 & , i = j \end{cases} \tag{4.9}$$

Integral equation (4.8) is Volterra equations of second kind and the integral is the convolution of $\psi_{i,m}^{(u)}(.)$ and $\chi_{m,j}^{(u)}(.)$; i.e., $\psi_{i,m}^{(u)} * \chi_{m,j}^{(u)}$. It gives the probability that user "u" starting in cell i, will be in cell j by t. The first part of the right-hand side of the equation is the probability that the user, being in cell i, never leaves cell i until the end of the period t. The second part of the right-hand side of the equation accounts for all cases in which the transition from i to j occurs via another cell $m \neq i$ applying the renewal argument. First, the probability of the user staying in cell i for a period of length τ and then going to cell m is given by $\psi_{i,m}^{(u)}(\tau)$. The handover to this new cell m can be interpreted as a renewal of the process because the expected behavior of the user from then on is the same irrespective of when the user enters cell m. Therefore, the probability of the user, who is in cell m at τ, being in cell j at t is given by $\chi_{m,j}^{(u)}(t-\tau)$. As the transition from i to m can occur at anytime between 0 and t, all possible transition times are considered by the integration over τ [61]. The numerical solution to solve evolution equation (4.8) is given by [62], and we implement the same approach. The evolution equation (4.8) can be re-written for discrete-time homogeneous semi-Markov process as follows:

$$\chi_{i,j}^{(u)}(k) = h_{i,j}^{(u)}(k) + \sum_{m=1}^{z}\sum_{\tau=1}^{k}\sigma_{i,m}^{(u)}(\tau)\chi_{m,j}^{(u)}(k-\tau) \tag{4.10}$$

where $h_{i,j}^{(u)}(k) = (1 - \Lambda_i^{(u)}(t))\delta_{i,j}$ and $\sigma_{i,m}^{(u)}(k) = \frac{d\psi_{i,m}^{(u)}(\tau)}{d\tau}$ can be approximated as follows, assuming that the time step is equal to the unit:

$$\sigma_{i,m}^{(u)}(k) = \begin{cases} \psi_{i,m}^{(u)}(1), & k = 1 \\ \psi_{i,m}^{(u)}(k) - \psi_{i,m}^{(u)}(k-1), & k > 1 \end{cases} \tag{4.11}$$

Since $\mathbf{P}^{(u)}$ is a right stochastic matrix, $\psi^{(u)}(k)$ and $\chi^{(u)}(k)$ will also be right stochastic matrices; i.e., $\sum_{j=1}^{z}\psi_{i,j}^{(u)}(k) = \sum_{j=1}^{z}\chi_{i,j}^{(u)}(k) = 1, \forall i,j \in \mathbb{C}$. The $\chi_{i,j}^{(u)}(k)$ indicates the probability of user "u" being in cell j after k amount of time from the time instant when he/she made the transition from somewhere to cell i. However, to predict the location of a user at every k' time steps, we have to estimate the probability $\hat{\chi}_{i,j}^{(u)}(k', s) = P(X_{s+k'}^{(u)} = j | X_0^{(u)} = i, t_{soj} = s)$, i.e., the probability of a user being in cell j after k' time, given that the current cell is i and the user has stayed in cell i for sojourn time $t_{soj} = s$. It can be evaluated as [58] below:

$$\hat{\chi}_{i,j}^{(u)}(k', s) = \frac{P(X_{s+k'}^{(u)} = j, t_{soj} = s, X_0^{(u)} = i)}{P(X_0^{(u)} = i, t_{soj} = s)} \tag{4.12}$$

$$= \frac{P(X_{s+k'}^{(u)} = j, t_{soj} = s | X_0^{(u)} = i)P(X_0^{(u)} = i)}{P(X_0^{(u)} = i, t_{soj} = s)} \tag{4.13}$$

$$= \frac{P(X_{s+k'}^{(u)} = j, t_{soj} = s | X_0^{(u)} = i)P(X_0^{(u)} = i)}{P(t_{soj} = s | X_0^{(u)} = i)P(X_0^{(u)} = i)} \tag{4.14}$$

$$= \frac{P(X_{s+k'}^{(u)} = j, t_{soj} = s | X_0^{(u)} = i)}{P(t_{soj} = s | X_0^{(u)} = i)} \tag{4.15}$$

$$= \frac{h_{i,j}^{(u)}(s+k') + \sum_{m=1}^{z}\sum_{\tau=s+1}^{s+k'}\sigma_{i,m}^{(u)}(\tau)\chi_{m,j}^{(u)}(s+k'-\tau)}{1 - \Lambda_i^{(u)}(s)} \tag{4.16}$$

Note that for $s = 0 : \hat{\chi}_{i,j}^{(u)}(k', s) = \chi_{i,j}^{(u)}(k)$. We will also leverage the steady-

state distribution of the semi-Markov model to analyze the long-term cell association of the users. This can help to identify the cells where users spend most of their time, and it can be further utilized to validate our proposed framework. The steady-state distribution of the semi-Markov model, i.e., $\Delta^{(u)} = [\Delta_1^{(u)}, \Delta_2^{(u)}, \Delta_3^{(u)}, ..., \Delta_z^{(u)}]$, is given as follows:

$$\Delta_j^{(u)} = \frac{\vartheta_j^{(u)} \xi_j^{(u)}}{\sum_{i=1}^{z} \vartheta_i^{(u)} \xi_i^{(u)}} \tag{4.17}$$

where $[\vartheta_1^{(u)}, \vartheta_2^{(u)}, \vartheta_3^{(u)}, ..., \vartheta_z^{(u)}]$ is a positive solution to the following balance equations:

$$\vartheta_j^{(u)} = \sum_{i=1}^{z} \vartheta_i^{(u)} p_{i,j}^{(u)},\ 1 \leq j \leq z \tag{4.18}$$

$$\sum_{i=1}^{z} \vartheta_i^{(u)} = 1 \tag{4.19}$$

and $\xi_j^{(u)}, 1 \leq j \leq z$ is the mean sojourn time of user "u" in cell j. Utilizing the past handover history of user "u" <time, Cell ID>, the probability transition matrix $\mathbf{P}^{(u)}$ and sojourn time distribution matrix $\mathbf{\Gamma}^{(u)}$ are initialized as follows [63]:

$$p_{i,j}^{(u)} = \frac{N_{i,j}^{(u)}}{N_i^{(u)}} \tag{4.20}$$

and

$$\Gamma_{i,j}^{(u)}(k) = \frac{N_{i,j,k}^{(u)}}{N_{i,j}^{(u)}} \tag{4.21}$$

where $N_{i,j}^{(u)}$ is the number of handovers of user "u" from cell i to j, $N_{i,j,k}^{(u)}$ is the number of handovers of user "u from cell i to j with a sojourn time less than or equal to k, and $N_i^{(u)}$ is the total number of handovers of user "u" from cell i. Whenever there is a handover from cell i to j, $p_{i,j}^{(u)}$ and $\Gamma_{i,j}^{(u)}(k)$ are updated and $\psi_{i,j}^{(u)}(k)$ is solved. Finally $\chi_{i,j}^{(u)}(k)$ and $\hat{\chi}_{i,j}^{(u)}(k', s)$ are computed. The cell with the highest probability is chosen as the predicted future destination, i.e., $\max_{j \in \mathbb{N}_i} \hat{\chi}_{i,j}^{(u)}(k', s)$ where $\mathbb{N}_i$ is the set of all neighboring cells of cell i. In this way, after every k' time steps, the next HO tuple information for each UE

$\{\mathbb{C}_N^u, \mathbb{T}_{HO}^u\}$ is generated wherein $\mathbb{C}_N^u$ is next probable cell of user "u" at time $\mathbb{T}_{HO}^u$.

4.3 Future Location Estimation

Let the UE's current location coordinates at time instant k be $l_k^u = (x_k^u, y_k^u)$ and the next cell HO tuple information for each UE be $\{\mathbb{C}_N^u, \mathbb{T}_{HO}^u\}$. The next task is to utilize this information for estimating the UE's future location coordinates in the next time step $k + k'$. Inspired by observation [64, 65] that users in a network usually move around a set of well-visited landmarks with landmark trajectory being fairly regular, we utilize the past mobility logs of UEs to estimate the most probable landmarks visited by each UE in each cell. This information is then utilized to estimate the direction of trajectory from the current location, while the distance to be travelled in that direction is estimated using the next cell HO time $\mathbb{T}_{HO}$. Let the coordinates of the most probable landmark for UE "u" in the next cell $\mathbb{C}_N^u$ be $l_{\mathbb{C}_N^u}^{LM} = (x_{\mathbb{C}_N^u}^{LM}, y_{\mathbb{C}_N^u}^{LM})$, then a unit vector $\hat{u}$ originating from the current coordinates in the direction of $(x_{\mathbb{C}_N^u}^{LM}, y_{\mathbb{C}_N^u}^{LM})$ is given as follows:

$$\hat{u} = \frac{[l_{\mathbb{C}_N^u}^{LM} - l_k^u]}{||(l_{\mathbb{C}_N^u}^{LM} - l_k^u)||} \tag{4.22}$$

where $||.||$ is the Euclidian norm operator. The future coordinates at time step $k + k'$ can be estimated as follows:

$$l_{k+k'}^u = l_k^u + \frac{\sqrt{(x_{\mathbb{C}_N^u}^{LM} - x_k^u)^2 + (y_{\mathbb{C}_N^u}^{LM} - y_k^u)^2}}{T_{HO}^u} * k' * \hat{u} \tag{4.23}$$

The pseudocode for the next location estimation algorithm is given in Algorithm 1 in appendix.

4.4 Experimental Evaluation

To realistically evaluate the proposed framework, we conducted an experimental study based on participatory sensing to analyze the applicability of the proposed model. In this experimental evaluation, the mobility pattern of a graduate student at the University of Oklahoma, Tulsa Campus, was logged for a month period in the Tulsa Campus region. The data gathered through the student's phone were used to build a semi-Markov model. This model was then used to predict his mobility pattern for the next whole week. The android application "LTE Discovery" was installed on the student's smartphone to continuously log the user's handover information. Once activated, the application continued to run in the background and updated the handover log whenever the user moved to some new cell. The logged information contains a time stamp and a new cell ID. In some places, such as indoor offices and cell overlapping regions, the test subject's equipment experienced a ping pong effect. The mobility history log was preprocessed to remove such entries, as has been done in [58], and only stable entries were utilized to build the semi-Markov model. Based on the recorded data set, four BSs were identified in the campus region, herein anonymously named A, B, C, and D. Two semi-Markov models were built (I and II) with time intervals of 1 hour and a quarter of an hour (15 minutes) respectively. A mobility pattern was predicted up to next 3-hour period. The sojourn time distribution matrix $\Gamma^{(u)}$ was computed for the test subject as done in [66]. Network scenario settings are listed in Table 4.1.

For prediction accuracy, each time the user entered a new cell, we calculated the probability of future locations for the next 3-hour period using the two semi-Markov models, and we compared it with actual mobility pattern. The

Table 4.1: Network scenario settings

No.	Parameter	Value
1	No. of Cells	4
2	Mean Sojourn Time (hours)	A: 0.33, B: 0.07, C: 2.95, D: 13.98
3	Speed (miles/hour)	max: 20
4	Prediction Interval (hour)	1/4, 1
5	Avg. no. of HOs (per day)	9
6	Area (sq. meters)	3000

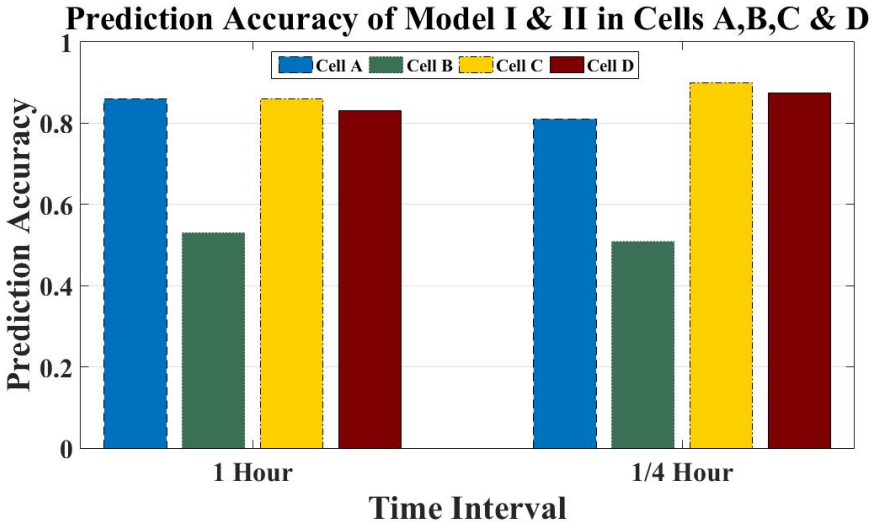


Fig. 4.3: Prediction accuracy

prediction accuracy results for each individual cell are presented in Fig. 4.3. As per the results, a minimum of approximately 50% and a maximum of 90% accuracy were achieved. The test subject had the least amount of sojourn time–around 2 minutes on average–in cell B, corresponding to the parking area, and it affected the training of sojourn time matrix; therefore, its prediction accuracy was the lowest. The user spent a relatively large amount of time in the rest of the cells, and the prediction accuracy was above 80% for all of the test cases. A smaller time interval effectively provided a better resolution; however it increased the operational complexity for the same prediction period, and the number of matrix multiplications increased. The difference in prediction accuracies between the two models having prediction time windows of 1 hour and a quarter of an hour, is not significant, at least for the scenario

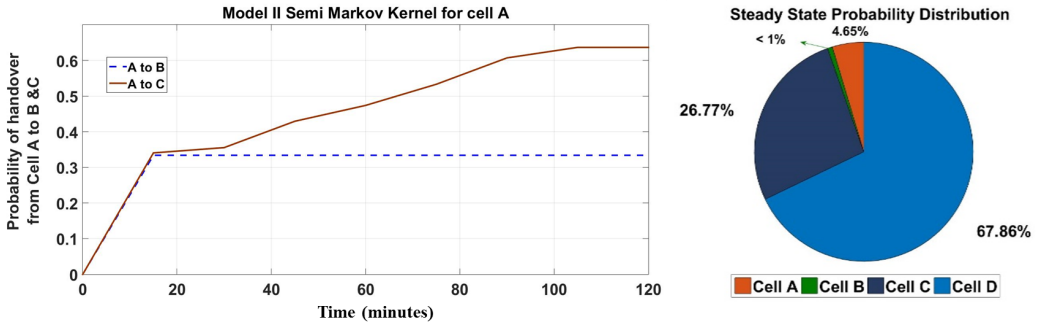


Fig. 4.4: (a) Semi-Markov kernel for Model II (b) Steady-state distribution

represented by our case study; therefore, the search for an optimal prediction window size can be avoided. Choosing the next two destinations with maximum probabilities instead of only one significantly increases the prediction accuracy–almost 100% in our test cases. However, this comes at the cost of decreased resource efficiency, as resources need to be reserved in more than one cell, and this factor amplifies in cases of incorrect predictions.

Model II's semi-Markov kernel for cell A, plotted in Fig. 4.4a, indicates the probability of transition of a user to neighboring cells B and C from cell A w.r.t time. From 0 to 15 minutes, the probabilities of transition to cells B and C are effectively the same, while from 15 minutes onwards, the probability of transition to cell C increases compared to cell B. This can be utilized to decide when, where, and for how long resources need to be reserved for each user for a successful and seamless handover between the cells. For instance, the necessary amount of resources could be initially reserved in both cells B and C during the first 15 minutes of transition to cell A. If the user stays in cell A for more than 15 minutes, then this could prompt the network to limit its resource reservation thereafter in cell C only, since that is the most likely handover to take place from the current cell.

The results for steady-state probability distribution are illustrated in Fig. 4.4(b). Accordingly, the user spends 67.86% of the time in cell D, followed by

26.77% in cell C, 4.65% in A and only 0.7% in B. The operator can utilize this information to identify the cells that are most likely to exhibit maximum traffic and plan the resources accordingly. For example, if the other users of the region exhibit a similar steady-state distribution to our test subject, then the network operator should have maximum capacity resource provision in cell D as compared to the other cells. It is important to highlight that the presented results are valid only for the considered network in which data are gathered. To be applicable to another network and set of users, the proposed mobility prediction model needs to be trained for that network and set of users.

4.5 Simulation Evaluation

We generated typical macro cell and SC-based network and UE distributions leveraging an LTE 3GPP standard compliant [67] network topology simulator in MATLAB. The simulation parameters' details are listed in Table 4.2. We used a wrap around model to simulate interference in an infinitely large network, thereby avoiding boundary effects. To model realistic networks, UEs were distributed non-uniformly in the coverage area such that a fraction of UEs were clustered around randomly located hotspots in each sector. Monte Carlo style simulation evaluations were used to estimate the average performance of the proposed framework. The real challenge here was the selection of a mobility trace generation model that realistically represents the behavior of actual cellular network users. Several such models have been proposed recently in literature, such as SLAW, SMOOTH, and Truncated Levy Walk [68]. Based on an extensive analysis of the pros and cons of these models, we chose the SLAW [69] mobility model. In contrast to conventional random walk models, where movement at each instant is completely random–chosen randomly

Table 4.2: Simulation settings

System Parameters	Values
Number of Base Stations	21
Number of UEs	84
Mobility model	SLAW
Prediction Interval	1 minute

from a set of allowed speeds and angles–SLAW has been shown to be a highly realistic mobility model. It exhibits all of the characteristics of real-world human mobility, i.e., (i) ***truncated power-law flights and pause-times***–the lengths of human flights that are defined to be straight line trips without directional change or pause have a truncated power-law distribution; (ii) ***heterogeneously bounded mobility areas***–people mostly move only within their own confined areas of mobility, and different people may have widely different mobility areas; (iii) ***truncated power-law inter-contact times***–the time elapsed between two successive contacts of the same persons follows truncated power-law distribution; and (iv) ***fractal waypoints***–people are always more attracted to more popular places. Therefore, the accuracy of the semi-Markov-based model tested using mobility traces generated by SLAW is very likely to represent its true performance in a real network. The SLAW mobility model was utilized to generate the HO traces of 84 mobile users for one week. Of that week, traces for the first six days were utilized to build and train the semi-Markov mobility model for each of the 84 UEs. Without loss of generality, and keeping operational complexity in mind, the prediction interval k' was set as 1 minute in our simulation study.

To benchmark the prediction accuracy of the semi-Markov-based model trained on six days training data, we utilized (4.10) and (4.16) to predict the serving cells of all UEs for the next whole day after every k' time step. At each time interval k, when the predicted future cell in the next time interval k' is the same as actual future cell, a score of 1 is given, otherwise it is 0. Accuracy is then calculated by summing the scores for all time instants and dividing that

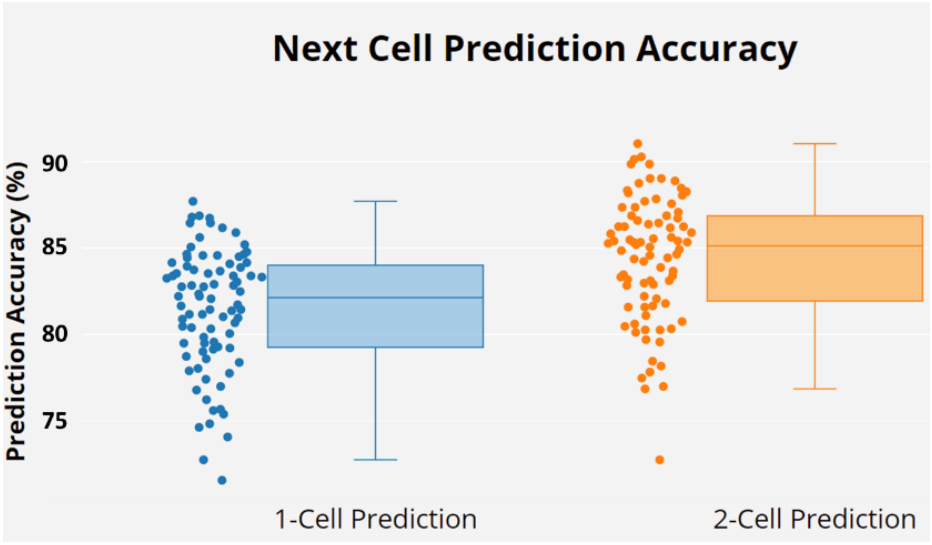


Fig. 4.5: Next cell prediction accuracy

total by the number of observations. The next cell prediction accuracy results are presented in Fig. 4.5. Accordingly, a maximum prediction accuracy of 87.70% was achieved, with a mean value of 81.46%, when choosing the top most probable cell among all future next cell candidates (1-Cell Prediction). The predictor performs exceptionally well, since the prediction interval is only 1 minute. This prediction can be further enhanced further by decreasing k' interval length. Fig. 4.6 illustrates that the mean prediction accuracy (denoted by dotted lines) monotonically decreases with an increase in k' interval's length. We could not decrease the prediction interval to less than 1 minute, as with the computational resources available for this study, the GA that is used to solve proactive SON functions in the upcoming chapters needed at least this minimum amount of time to find a feasible solution. However, it is anticipated that if more powerful computational resources are leveraged to reduce the convergence time of the GA, then better mobility prediction accuracy may be achieved. We also analyzed the effect of choosing two of the most-probable future next cell candidates (2-Cell Prediction) instead of one. The prediction accuracy received a slight boost, with the mean value reaching

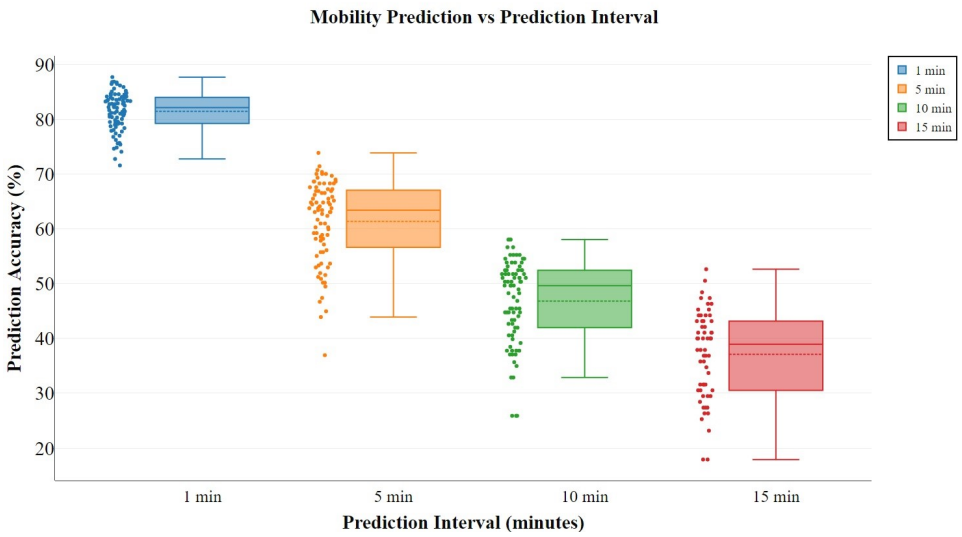


Fig. 4.6: Effect of prediction interval on next cell prediction accuracy

up-to 84.39%. However, this gain is not significant given that it already has high accuracy.

Next, based on the next cell HO tuple information for each UE $\{\mathbb{C}_N^u, \mathbb{T}_{HO}^u\}$, we compared the actual and predicted number of UEs per cell. Let $|\mathbb{U}_j(t+1)|$ be the number of users predicted to be in cell j at time $t+1$. This consists of users who (i) just entered into cell i at time t and will be in cell j at time $t+1$ given by the following equation:

$$\mathbb{U}_j(t+1) := \{\forall u \in \mathbb{U} | j = arg\,max_{m \in \mathbb{C}}(\chi_{i,m}^{(u)}(k=1))\} \tag{4.24}$$

and (ii) users who are in cell i and have stayed in cell i for sojourn time $t_{soj} = s$ and will be in cell j at time $t+1$ given by the following equation:

$$\mathbb{U}'_j(t+1) := \{\forall u \in \mathbb{U} | j = arg\,max_{m \in \mathbb{C}}(\hat{\chi}_{i,m}^{(u)}(k'=1,s))\} \tag{4.25}$$

Therefore, the total number of UEs predicted to be in cell j at time $t+1$ will be as follows:

$$|\mathbb{U}_j(t+1)| = |\mathbb{U}_j(t+1)| + |\mathbb{U}'_j(t+1)| \tag{4.26}$$

As evident in the Fig. 4.7, the mobility prediction model is able to predict

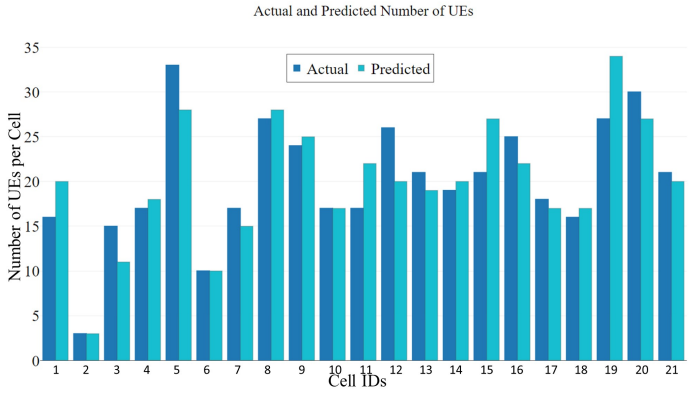


Fig. 4.7: Actual and predicted number of UEs per cell

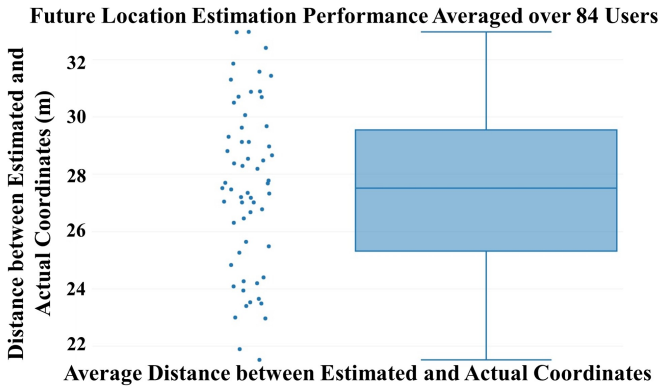


Fig. 4.8: Future location coordinates estimation performance

the number of UEs in most of the cells at the next time interval with high accuracy. Next, based on the next cell HO tuple information for each UE $\{\mathbb{C}_N^u, \mathbb{T}_{HO}^u\}$, future location coordinates were estimated using Algorithm 1 for all UEs for a 1 hour simulation duration after every k' time steps. The average estimation performance is illustrated in Fig. 4.8, according to which the maximum distance error between the estimated and actual coordinates was around 33 meters, with a mean value of around 27.5 meters. The location estimation algorithm performed exceptionally well. One particular reason for the high accuracy is that the SLAW model is for pedestrian users. Therefore, the location of a user changes slowly as a function of time and thus remains relatively more predictable. With high speed, accuracy is expected to degrade; however,

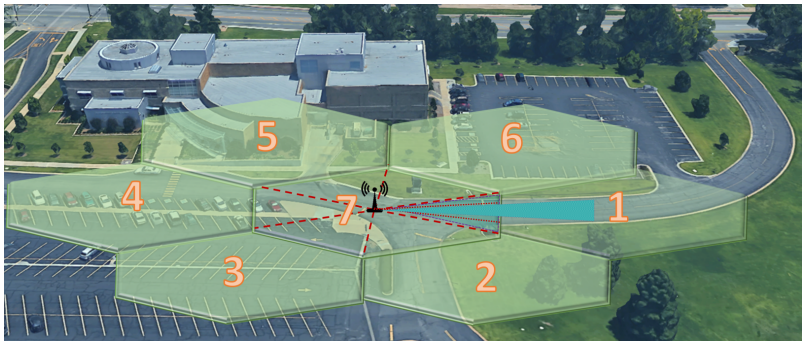


Fig. 4.9: Leveraging geographical knowledge for facilitating user/cell discovery

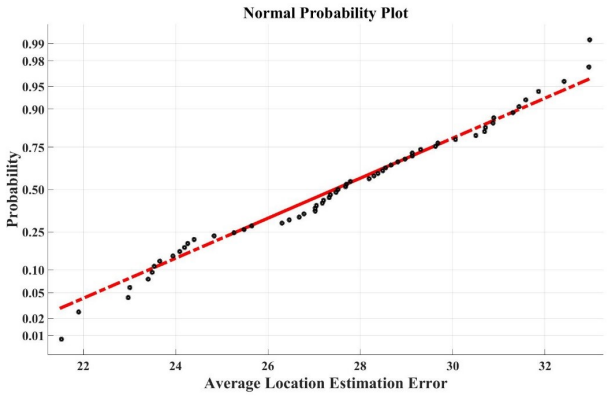


Fig. 4.10: Normal probability plot for average location estimation error

knowledge of the street/road layout can be exploited to maintain accuracy. This idea is illustrated in Fig. 4.9, where the HO record for a user's transition from cell 1 to cell 7 is superposed onto a local map. With this superposition, the spatial resolution of the HO prediction can be narrowed down from the whole cell boundary to a narrow track, and the temporal prediction can also be improved by incorporating the typical speed limit of the track and the RSRP gradient into the prediction time produced by the mobility prediction model for that user. However, this is beyond scope of this contribution and will be the subject of a future study. The normal probability plot for the average location estimation error is depicted in Fig. 4.10, that is basically a plot of the ordered observations from a sample against the corresponding percentage points from the standard normal distribution. If the data come

from a normal distribution, then they will fall on an approximately straight line. As per the figure, normal distribution can be good approximation of the average location estimation error distribution.

4.6 Gain of the semi-Markov-based Mobility Prediction Framework

Mobility prediction accuracy is the parameter investigated in this chapter as the holistic performance of a proactive-SON-enabled cellular system depends on how accurate its predictions are. The gain of the semi-Markov-based mobility prediction framework (SMPF)-based cellular system can be evaluated as follows [70]:

$$Gain = \frac{\lambda_{np} - \lambda_{smpf}}{\lambda_{np}} \tag{4.27}$$

where λ_{np} is the resource utilization cost in a conventional non-predictive CS, and λ_{smpf} is the expected resource utilization cost for an SMPF-enabled CS given as follows:

$$\lambda_{smpf} = \alpha_p(RUC_c) + (1 - \alpha_p)(RUC_{ic}) \tag{4.28}$$

Here, α_p is the prediction accuracy, and RUC_c and RUC_{ic} are the resource utilization costs for correct and incorrect predictions respectively. These can be handover resource reservation costs, resource block reservations for capacity, caching and waking up next BS, among other things. An incorrect prediction may degrade the overall system performance, since it reserves resources that could otherwise be used for other users. The gains for different RUCs with prediction accuracy are plotted in Fig. 4.11. A fixed value of 200 is considered for λ_{np}. In (a), when RUC_c is half of a non-predictive CS (λ_{np}),

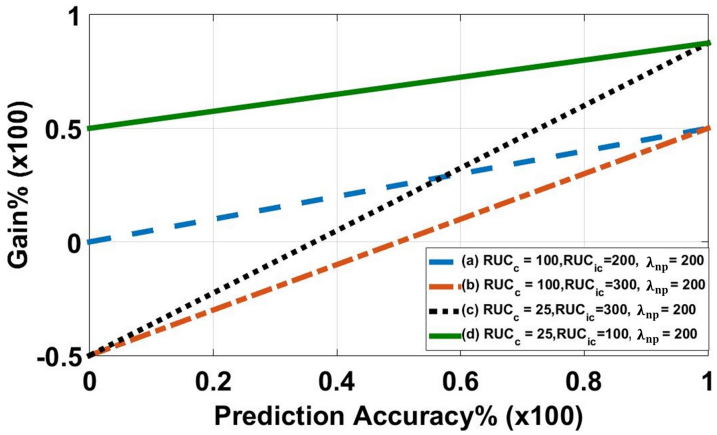


Fig. 4.11: Gain of SMPF vs. prediction accuracy

and RUC_{ic} is the same as λ_{np}, then the gain is always positive and lies in the range of (0% to 50%). When RUC_{ic} is increased to 300 in (b), then the gain can be negative, and at 50% prediction accuracy, we obtain a gain of 0% (the same performance as that of a non-predictive CN). When RUC_c is decreased to 25 in (c), then the gain achieved rises to the maximum. When RUC_{ic} is also decreased to half of λ_{np} as in (d), then the gain is always positive and $\geq 50\%$ for all prediction accuracies. While the gain is a generic measure, and the evaluation of specific values is beyond the scope of this contribution, it provides a framework for assessing the gain of the SMPF and its minimum accuracy needed to achieve any gain.

4.7 Conclusion

This chapter proposed a novel spatiotemporal mobility prediction model employing the innovative concept of estimating future user locations that in turn can empower SON functions like ES, MLB, CCO and MRO. Experimental and simulation evaluations demonstrated that the proposed model achieved a high prediction accuracy of above 80% for the majority of the cells. In next two chapters, the gain of the proposed mobility prediction framework will be

evaluated for different SON use cases.

CHAPTER 5

Mobility Prediction-Based, Autonomous, Proactive Energy Saving (AURORA) Framework for Emerging Ultra-Dense Networks

> The best way to predict the future is to create it.
>
> Abraham Lincoln

Increased network-wide energy consumption is a paramount challenge that hinders wide-scale UDNs deployments. While several ES enhancement schemes have recently been proposed, these schemes have one common tendency: they operate in reactive mode; i.e., to increase ES, cells are switched ON/OFF reactively in response to changing cell loads. Although, significant ES gains have been reported for such ON/OFF schemes, the inherent reactiveness of these ES schemes limits their ability to meet the extremely low latency and high QoS expected from future cellular networks vis-à-vis 5G and beyond. To address this challenge, in this contribution we propose a novel user mobility prediction-based, autonomous, proactive, ES (AURORA) framework for future UDNs. Instead of passively observing changes in cell loads and then reacting to them, AURORA uses past HO traces to determine future cell loads. This prediction is then used to proactively schedule SC sleep cycles. AURORA also incorporates the effect of CIOs for balancing the load between cells to ensure QoS while maximizing ES. Extensive system-level simulations, leveraging realistic SLAW model-based mobility traces, demonstrate that AURORA can achieve a significant energy reduction gain without a noticeable impact on QoS.

5.1 Introduction

The current exponential mobile data traffic escalation is a precursor of an imminent "capacity crunch". Against backdrop, extreme network densification through the deployment of a large number of SCs has emerged as the most yielding solution to achieve the 1,000-fold capacity gain goal [5]. However, the ultra-dense deployments of SCs is on a direct collision path with the economically viable and energy efficient deployment vision of 5G. This is due to the high aggregated network energy that "always ON" SCs are bound to consume in a UDN. In addition to a higher carbon footprint, this translates into higher OPEX. Although SCs have a relatively low power consumption profile, the always ON approach increases overall network-wide energy consumption [71]. This is because the load-independent power consumption (circuit power) component in SCs constitutes a much larger portion of over-all power consumption [72]. As a result, with the advent of UDNs, the need for ES schemes will be even more compelling. The consensus among the research community is that to avert a possible energy crunch in 5G and to achieve economic viability, the $1,000\times$ capacity increase must be achieved at a power consumption that is similar to or lower than that of legacy networks [73].

5.1.1 Related Work

Energy consumption in cellular systems can be reduced significantly by turning OFF underutilized cells during off-peak hours or by optimizing resource allocation such that minimum energy is consumed per bit transmission [73, 74, 75, 76]. To exploit these approaches, 3GPP has recently adopted ES as a key SON function [77], and ES has been extensively studied in literature. Energy-saving enhancement, with a focus on optimizing resource allocation despite its relatively small gain compared to turning ON/OFF under-utilized

BSs, has been studied more extensively compared to later approach [73]. Resource allocation optimization can reduce the energy consumption to only a limited degree for a given system throughput target. The ES of cellular systems can be further enhanced significantly by switching under-utilized BSs to sleep mode or turning them OFF entirely during off-peak time [74, 75, 76, 78]. In this direction of research, some recent works demonstrate promising results in terms of potential ESs [79, 80, 81, 82, 83, 84, 85, 86, 87, 88, 89, 90, 91, 92]. However, to the best of our knowledge, existing ES approaches fall short of the mark for 5G requirements due to the following four limitations:

1. **Reactive mode of operation**. Conventional ES SON algorithms are designed to switch OFF/ON cells after detecting network conditions that have already taken effect. For example, when congestion is detected in a network, a non-convex, NP-hard ES algorithm is usually solved to identify certain sleeping/OFF cells that should be switched ON, or using the same process, certain cells are switched OFF when a low load is observed in specific cells. This is an improvement on fixed-timer based switching ON/OFF [93] which can, at best, follow a coarse, statistical, spatio-temporal traffic pattern and thus achieves ES at the cost of QoS. However, given the acute dynamics of the traffic and cellular environment, by the time congestion or low traffic conditions are detected, and a realistic non-convex, NP-hard ES algorithm is solved to produce a new network ON/OFF configuration that is optimal for the observed network conditions, the conditions may have already changed. Therefore, the newly determined switch ON/OFF vector is likely to be suboptimal before it can be actuated. This problem can be exacerbated particularly in 5G, where a motely of traffic and a plethora of cell types mean that the dynamics of a cellular eco-system will be even more swift.

2. **Difficulty in meeting 5G low latency.** Base stations require a certain amount of time to wake up from a sleep cycle [94]. For a user entering a sleeping cell, this time to wake up will add to the latency experienced by the user. This demands a paradigm shift from the conventional reactive design of ES algorithms towards proactive characteristics to cope with the extremely low latency requirements of 5G in a more agile fashion.

3. **Impractical cell discovery.** The following is a key challenge in switching OFF-based ES schemes: how to discover an OFF cell when users enter into the physical coverage area of the OFF cell. Existing ES schemes either overlook this challenge, or propose solutions that exploit neighboring cells or a master controller to wake up the cell when enough users enter into the coverage area of the OFF cell. While this approach may work in a low user density network with large macro cells with relatively less stringent QoS requirements, such as LTE, it may not scale to 5G because of signaling overhead, delays, and the cost of missing out OFF SCs for off-loading.

4. **Self-organizing networks' conflict prone design.** The other caveat with conventional ES solutions is that they are oblivious to the fact that multiple SON functions may be prone to hidden or undesired conflict when implemented together in a network [95, 5]. Two SON use cases that become highly relevant to the ES in HetNets are CCO and LB [77] because of the overlap in their optimization parameter set: transmission power and CIOs. When ES switches OFF some cells, it may force some users to be associated with neighboring ON cells and overload them, thereby conflicting with the CCO and LB SON functions. As explicated in [95], such a conflict prone ES solution design can actually degrade a network's performance instead of improving it.

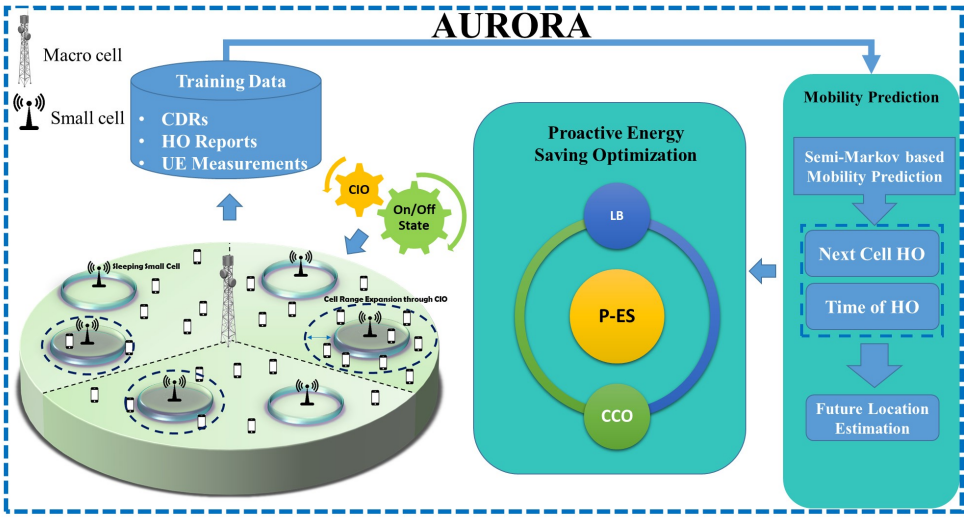


Fig. 5.1: The AURORA framework

5.2 The AURORA Framework

To address the aforementioned limitations, we propose the AURORA framework (Fig. 5.1) by building on the lines of a big data empowered SON framework [5]. The key idea is to make emerging cellular systems artificially intelligent and autonomous, so that they can anticipate user mobility behavior. This intelligence is then used to formulate a novel ES optimization problem that proactively schedules SC sleep cycles to divert and focus the right amount of resources when and where needed while satisfying QoS requirements.

In this section we present the analytical model development of the AURORA framework, whose three corner stones are as follows:

- The semi-Markov-based spatiotemporal next cell prediction, presented in chapter 4.
- The mapping of next cell prediction to future user location estimation, presented in chapter 4.
- Proactive-ES optimization based on future user location estimation.

5.2.1 Network Model and Assumptions

The AURORA framework proposed in this contribution only focuses on the downlink of cellular systems for the sake of conciseness. It is assumed that all mobile devices and SCs have omnidirectional antennas with a constant gain in all directions, while macro cells have directional antennas. A frequency reuse of 1 is considered, and the same band is utilized by the macrocell and the SCs. A full buffer traffic model is used for each user; i.e., there are always data available to be sent for a user with a constant bit rate (CBR) service. A centralized C-SON architecture is assumed wherein a centralized server in the core network performs a system-wide proactive-ES optimization. Moreover, HO traces that include the location-stamped information of past cell transitions, such as cell IDs, RSRPs, and call detail records are assumed to be available to the C-SON server.

5.2.2 Proactive ES Optimization

Given the next probable HO tuple and estimated future location $l^u_{k+k'}$ for all users determined through the semi-Markov model presented in chapter 4, we devise an ON-OFF sleeping mechanism for SCs for the next time step $k+k'$ to minimize network-wide energy consumption. The sleeping schedule is ensured to satisfy the coverage KPI and QoS requirements of each UE located at its estimated future location $l^u_{k+k'}$ as well as to satisfy the maximum loading constraint for each BS. The total instantaneous power consumption of a cell can be given by the sum of the circuit and the transmit power as follows [72]:

$$P_c^{\text{total}} = \pi^c (P_{CT}^{\text{c}} + \eta_c . P_t^{\text{c}}) \tag{5.1}$$

where P_{CT}^{c} is the constant circuit power, which is drawn if a BS in cell c is active and is significantly reduced if the BS goes into sleep mode; P_t^{c} is the

transmit power of cell c; η_c denotes the load; and π^c is an indicator variable that will be 1(0) for an ON(OFF) BS in cell c. One way in which to quantify ESs is to leverage the PM criterion of the energy consumption ratio (ECR) [96, 97]. This ECR for a cell is defined as the amount of energy consumed in Joules per bit of information that is reliably transmitted in that cell, calculated as follows:

$$ECR_c = \frac{P_c^{\text{total}}}{\sum_{\mathbb{U}_c} \omega_B^u * f(\gamma_u^c)} (Joules/bit) \tag{5.2}$$

where $f(\gamma_u^c)$ is a function that returns the achievable spectral efficiency of user "u" at a given SINR γ_u^c and ω_B^u is the bandwidth assigned to user "u". The $f(\gamma_u^c)$ can be defined to take into account post-processing diversity gains such as the ones harnessed by MIMO and/or the loss incurred by system-specific overheads using $f(\gamma_u^c) := A \log_2(1 + B(\gamma_u^c))$. Here, A and B are constants taken as 1 in our simulation studies without a loss of generality. The SINR $\hat{\gamma}_u^c$ at an estimated user location $l_{k+k'}^u$ at time step $k + k'$ when associated with a cell c is defined as the ratio of reference signal received power $P_{r,u}^c$ by user "u" from cell c to the sum of the reference signal received power by user "u" from all cells i such that $\forall i \in \mathbb{C}/c$, and the noise variable κ:

$$\hat{\gamma}_u^c(k + k') = \left[\frac{P_t^c G_u G_u^c \delta \alpha (d_u^c)^{-\beta}}{\kappa + \sum_{\forall i \in \mathbb{C}/c} P_t^i G_u G_u^i \delta \alpha (d_u^i)^{-\beta}} \right]_{k+k'} \tag{5.3}$$

where P_t^c is the transmit power of cell c; G_u is the gain of user equipment; G_u^c is the gain of the transmitter antenna of the cell c, as seen by the user "u;" δ is the shadowing observed by the signal; α is the path loss constant; d_u^c represents the distance of the estimated location of user "u;" i.e., $l_{k+k'}^u$ from cell c and β is the path loss exponent. The time subscript on the right hand side of (5.3) and in rest of the chapter indicates that all terms enclosed within $[.]_{k+k'}$ are considered for the next time step $k+k'$. Within the scope of

this contribution, it is assumed that shadowing estimate information for the estimated user location is available with a normally distributed error. In a practical network, both channel maps that build on the MDT reports and the collected channel quality indicator reports can be utilized to estimate channel gains in estimated locations. This $\hat{\gamma}_u^c(k+k')$ is a fully loaded SINR expression and is valid only when all cells are fully utilized. The actual interference from neighboring cells based on their respective loads is utilized as follows to calculate the SINR for data transmission:

$$\gamma_u^c(k+k') = \left[\frac{P_t^c G_u G_u^c \delta \alpha (d_u^c)^{-\beta}}{\kappa + \sum_{\forall i \in \mathbb{C}/c} \eta_i P_t^i G_u G_u^i \delta \alpha (d_u^i)^{-\beta}} \right]_{k+k'} \tag{5.4}$$

where η_i denotes the cell load in a cell i at time step $k+k'$. This method of weighting the interference power received from each cell with its current resource utilization results in a certain coupling of the total interference with different cell utilizations. More loaded cells contribute more interference power than less loaded ones [98]. For an LTE network, an instantaneous cell load can be defined as the ratio of PRBs occupied in a cell during a transmission time interval (TTI) to the total PRBs available in the cell. This indicator is available as a standard measurement in LTE as "UL/DL total PRB usage." The number of PRBs allocated to each user depends on the QoS that the user requires and the achievable SINR. For instance, if the QoS is defined in terms of the required data rate, more PRBs are assigned to a user with a higher rate requirement and/or one with a lower SINR. The total load of cell c at time step $k+k'$ will be the fraction of the total resources in the cell needed to achieve the required rate of all users of a cell given as follows:

$$\eta_c(k+k') = \left[\frac{1}{N_c} \sum_{\mathbb{U}_c} \frac{\hat{\tau}_u}{\omega_B \log_2(1+\gamma_u^c)} \right]_{k+k'} \tag{5.5}$$

where ω_B is the bandwidth of one resource block, N_c is the total number of

resource blocks in cell c, $\hat{\tau}_u$ is the minimum required rate of the user, and $\mathbb{U}_c$ is the number of active users connected to a cell c. It is a virtual load, as it is allowed to exceed 1 to provide us with a clear indication of how overloaded a cell is. The required rate in the numerator is the minimum bit rate required by the user, depending on the QoS requirements of the services and user subscription level. In current LTE standards, an exact method does not exist to estimate the throughput required by the user. Only the historical throughput of a user can be estimated after the allocation of resources. However, 3GPP standards define a metric called the QoS class identifier (QCI). The primary purpose of the QCI is to prioritize users based on their required resource type, packet delay susceptibility and packet error loss rate. The definition of the desired throughput can build on the QCI. In a more robust approach that leverages network analytics, $\hat{\tau}_u$ can be modeled as a function of subscriber behavior, subscription level, service request patterns, as well as the applications being used [5]. The set of users connected to cell c is determined by the following user association criterion:

$$\mathbb{U}_j := \{\forall u \in \mathbb{U} \mid j = \arg\max_{\forall c \in \mathbb{C}} (P^c_{r,u_{dBm}} + P^c_{CIO_{dB}})\} \tag{5.6}$$

where $P^c_{r,u_{dBm}}$ is the true reference signal power in dBm received by user "u" from cell c, and $P^c_{CIO_{dB}}$ is the bias parameter (the CIO). The term CIO is a common identifier for, e.g., the real Qrxlevminoffset and Qqualminoffset parameters for cell selection, the Qoffset parameter in cell reselection, and the Oxx parameters for event Ax measurements, as preparation for HO procedures in radio resource management [99]. This CIO is primarily used to offset the lower transmit power of SCs to transfer a higher load to them (Fig. 5.2). In case some underutilized cells are turned OFF, the remaining cells need to have maximum utilization to cater for the transferred load from underutilized

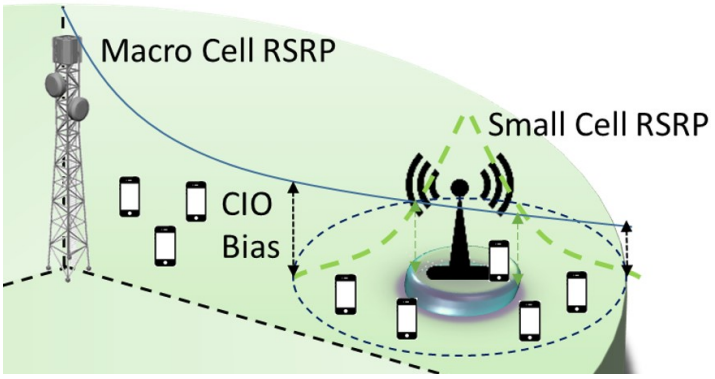


Fig. 5.2: CIO bias

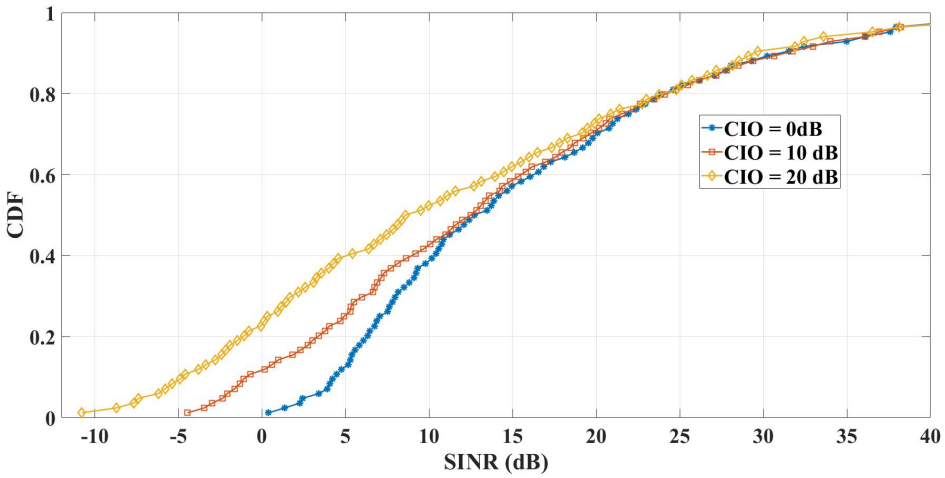


Fig. 5.3: Average UE SINR (dB) vs. CIOs

cells. However the downside of biasing is that UEs are no longer necessarily connected to the strongest cell. As a result, the SINR is bound to be lower with higher CIO values, as illustrated in Fig. 5.3. However, CIO is still a necessary measure to balance the loads. The capacity loss due to a drop in SINR can partially be offset if the serving cell has more free PRBs that can be allocated to that user, compared to PRBs in the previous serving cell, to satisfy the required QoS. This highlights the importance of the CIO parameter as a knob to control the trade-off between network LB, CCO, and energy consumption.

It is important to highlight here that in case of ES optimization with guaran-

teed minimum QoS requirements, it does not make sense to look at throughputs, since the UEs either receive an exact constant bit-rate or they are unsatisfied. Therefore, a more appropriate PM to analyze is the number of unsatisfied or dropped users "N_{us}," given as follows [48]:

$$N_{us}(k+k') = \left[\sum_{c} \max(0, \sum_{\mathbb{U}_c} 1.(1 - \frac{1}{\eta_c}))\right]_{k+k'} \tag{5.7}$$

where $\sum_{\mathbb{U}_c} 1.$ is the total number of users in cell c, while $(1-\frac{1}{\eta_c})$ is a modulation parameter that indicates what percentage of users in that cell are unsatisfied. Here, by definition from (5.5), η_c is allowed to exceed 1 to provide a clear indication of how overloaded a cell is. When $\eta_c = 1$, the inner summation in (5.7) will be 0, meaning that all users in cell c are satisfied. When $\eta_c = 2$, the inner summation will be equal to half of the number of users of cell c, meaning that half of the users are satisfied. The outer summation is the total number of unsatisfied users in the whole network while a max operator is used, since the number of unsatisfied users cannot be negative in under-loaded cells. Unsatisfied users would not be allowed to enter the system, or they would be dropped if they are already active.

Now we formulate the general energy consumption minimization problem for time step $k + k'$ as (5.8-5.10e):

$$\min_{\boldsymbol{\pi^c}, \boldsymbol{P^c_{CIO}}} \sum_{\mathbb{C}} [ECR_c]_{k+k'} \tag{5.8}$$

$$\min_{\boldsymbol{\pi^c}, \boldsymbol{P^c_{CIO}}} \sum_{\mathbb{C}} \left[\frac{\pi^c(P^c_{CT} + \eta_c.P^c_t)}{\sum_{\mathbb{U}_c} \omega^c_u \log_2(1 + (\frac{P^c_t G_u G^c_u \delta \alpha (d^c_u)^{-\beta}}{\kappa + \sum_{\forall i \in \mathbb{C}/c} \eta_i P^i_t G_u G^i_u \delta \alpha (d^i_u)^{-\beta}}))}\right]_{k+k'} \tag{5.9}$$

where $\mathbb{U}_j := \{\forall u \in \mathbb{U} \mid j = \arg\max_{\forall c \in \mathbb{C}} (P^c_{r,u_{dBm}} + P^c_{CIO_{dB}})\}$

$$P^c_{CIO.min} \leq P^c_{CIO} \leq P^c_{CIO.max} \forall c \in \mathbb{SC} \tag{5.10a}$$

$$\pi^c \in \{0,1\} \forall c \in \mathbb{SC} \tag{5.10b}$$

$$\frac{1}{|\mathbb{C}|} \sum_{\mathbb{C}} \frac{1}{|\mathbb{U}_c|} \sum_{\mathbb{U}_c} 1(P^c_{r,u} \geq P^c_{th}) \geq \bar{\omega} \tag{5.10c}$$

$$\tau_u \geq \hat{\tau}_u \forall u \in \mathbb{U} \tag{5.10d}$$

$$\eta_c \leq \eta_T \forall c \in \mathbb{C} \tag{5.10e}$$

The objective is to optimize the parameters $\boldsymbol{\pi^c}, \boldsymbol{P^c_{CIO}}$ of SCs ($\mathbb{SC}$) such that the energy consumption ratio in all cells is minimized while ensuring coverage reliability and the satisfaction of user throughput requirements. The first two constraints define the limits for the CIOs and ON/OFF state array respectively. These are the constraints that will determine the size of the solution search space. The third constraint is to ensure minimum coverage. Here, P^c_{th} is the threshold for the minimum received power for a user to be considered covered, $\bar{\omega}$ defines the area coverage probability (a QoS KPI) that an operator wants to maintain, and $1(.)$ denotes an indicator function. The fourth constraint ensures that each user receives the required minimum bit rate, depending on the QoS requirements of the service and the user's subscription level. This is due to the fact that to achieve the ECR minimization objective, the CIO of the remaining ON SCs may be increased to offload users of switched OFF cells into their coverage umbrella. The consequences are that the received power $P^c_{r,u}$ of offloaded users may become worse, leading to degraded SINR and throughputs. The effect of decreased SINR can be offset by allocating more resources only if the received power by the user is above a certain threshold. Therefore, this fourth constraint ensures that the minimum throughput is guaranteed for all users in all cases. However, this can only oc-

cur when the number of resources available in a cell is sufficient to meet user requirements; therefore, this constraint is complemented by a constraint on cell load $\eta_c \leq \eta_T$ (load threshold) with $\eta_T \in (0, 1]$.

The formulated combinatorial optimization problem in (5.8-5.10e) contains both continuous $\boldsymbol{P^c_{CIO}}$ and binary $\boldsymbol{\pi^c}$ decision variables. It can be identified as a mixed-integer non-linear programming problem (MINLP). The inherent coupling of the ON/OFF state vector, CIOs, and cell loads indicate that it is a large-scale non-convex optimization problem. As we are dealing with two problem parameters per cell whose effects on the optimization function are not independent, the complexity is expected to grow exponentially with the number of cells. Therefore, an exhaustive search for the optimal parameters may not be practical for a large-size network due to a high complexity time search that needs to be done in real time. For a practical scenario, with 50 SCs and only CIO as the optimization variable with 10 possible values available at each SC, we already have 10^{50} possible settings. This is approximately equal to the number of atoms on earth. Therefore, to solve the formulated ES problem, we utilized GA [100] with the pseudocode given in appendix since it is considered attractive heuristic technique for a multi-variable MINLP problem with a large variable count and enormous search space. Genetic algorithms are collectively a class of artificial intelligence algorithms based on a natural selection process that mimics biological evolution. In contrast to classical optimization wherein a single point is generated at each iteration, and the sequence of points gradually approaches an optimal solution, GAs generate a population of points at each iteration, and the best point in the population approaches an optimal solution. Due to its random nature, a GA significantly improves the chances of finding a global solution, especially for highly non-linear objective functions. It is also important to note that the

GA starts from a random parameter set in the solution space; therefore, it does not require a feasible point to start a search. Based on the estimated network state for time step $k + k'$, the AURORA framework consequently devises the optimal ON/OFF state array and CIO values for all the SCs ahead of time such that the energy consumption ratio of the whole network is minimized. The ON/OFF state array and CIO values remain fixed from k to k'. In a practical network, SCs need some non-zero time to switch their states; therefore, the proposed strategy allows ample time of k' duration for SCs to switch to an optimal ON/OFF state.

5.3 Performance Analysis

In this section, we analyze the potential ESs resulting from the application of the AURORA framework on HetNets. We have benchmarked its performance against four schemes; (i) **n**e**a**r-Optimal pe**r**formance bou**n**d (NARN) wherein it is assumed that AURORA estimates the future location and channel estimate at that location with 100% accuracy; (ii) **all** cells **ON** with **hom**ogeneous **net**work settings (AllOn-HomNet) wherein all cells are ON, and no CIO is utilized for SCs; (iii) **all** cell **on** with **het**erogeneous **net**work settings (AllOn-HetNet) wherein all cells are ON, and a fixed CIO of 10 dB is utilized for all SCs; and (iv) a reactive scheme that is simulated by delaying user location information, i.e., optimization with $\eta_T = 1$ is done based on the location information of the previous 1 minute.

5.3.1 Simulation Settings

We generated typical macro cell and SC-based network and UE distributions leveraging an LTE 3GPP standard compliant [67] network topology simulator in MATLAB. The simulation parameters' details are listed in Table 5.1. We used a wrap around model to simulate interference in an infinitely large net-

Table 5.1: Network simulation settings

System Parameters	Values
Number of Macro Base Stations	7 with 3 Sectors per Base Station
Small Cells per Sector	5
Number of UEs	Mobile: 84, Stationary: 336
LTE System Parameters	Frequency = 2 GHz, Bandwidth = 10 MHz
Macro Cell Tx Parameters	Tx Power = 46 dBm, Tilt = 102^0
Small Cell Tx Parameters	Tx Power = 30 dBm, CIO = 0 to 10 dB
Base Station Heights	Macro BS = 25m, Small BS = 10m
Area Coverage Probability	100%
Total Simulation Duration	1 hour

work, thereby avoiding boundary effects. To model realistic networks, UEs were distributed non-uniformly in the coverage area such that a fraction of UEs were clustered around randomly located hotspots in each sector. Monte Carlo-style simulation evaluations were used to estimate the average performance of the proposed framework. Furthermore, SLAW [69] was used as the mobility model to generate HO traces of 84 mobile users for 1 week. Of this week, traces for the first six days were utilized to build and train the semi-Markov mobility model for each of the 84 UEs. Moreover, an additional 336 stationary UEs (80% of the total UEs [101]) were deployed to generate additional loading on the network. For traffic demand, we considered two scenarios (i) ***low traffic demand*** comprising of five different uniformly distributed UE traffic requirement profiles corresponding to desired throughputs of 24 kbps (voice), 56 kbps (Text Browsing), 128 kbps (Image Browsing), 512 kbps (FTP) and 1,024 kbps (video), and (ii) ***high traffic demand*** wherein all UEs are video users. Without a loss of generality and keeping operational complexity in mind, the prediction interval k' was set as 1 minute in our simulation study.

5.3.2 Quantifying the ES Potential of the AURORA Framework

The ECR of AURORA and NARN for low and high traffic demands with varying values of load thresholds η_T along with that of AllOn-HomNet, AllOn-HetNet and state-of-the-art reactive schemes averaged over a 1-hour duration are visualized in Fig. 5.4. Note that to visualize the ECR ranges for both

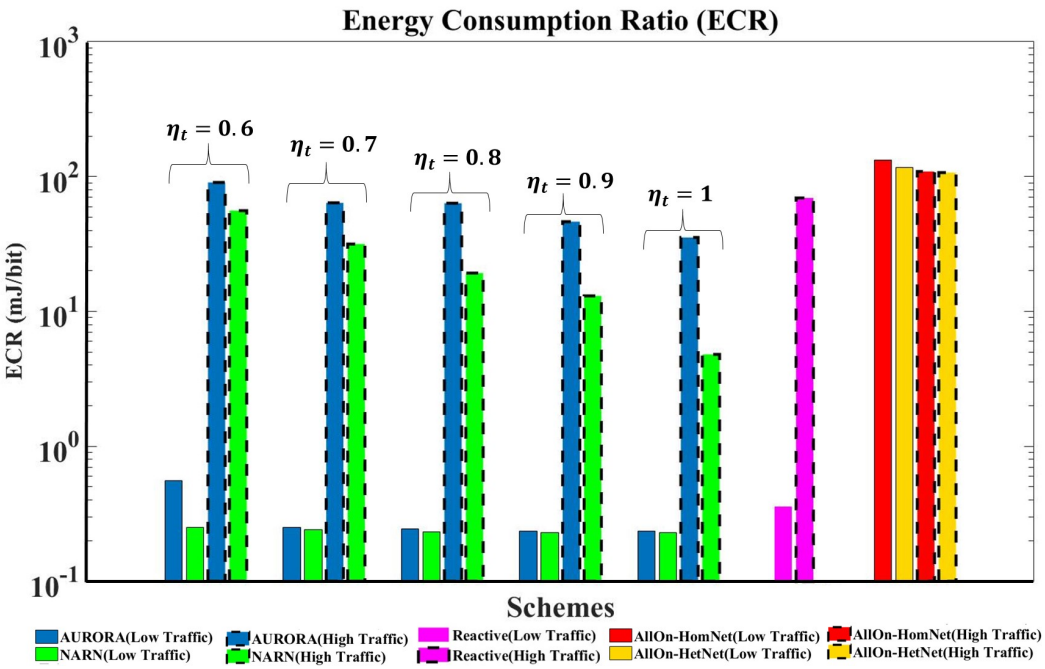


Fig. 5.4: Energy consumption ratio (ECR)

traffic classes in same figure, the y-axis has been plotted in a logarithmic scale. The load threshold range is [0.6, 1], since the P-ES optimization algorithm (5.8) returned no feasible point below 0.6. It was observed that the ECR values are higher for a high traffic demand scenario, as a higher number of SCs need to be switched ON to cater for the high load. Moreover AURORA exhibited a linearly decreasing trend with increasing values of η_T. It is significantly less than the conventional AllOn schemes for all load threshold values. The reason is that for AllOn schemes, all cells are ON at all times, this increases energy consumption which is bound to further escalate with densification. At lower η_T values, the ECR for AURORA is higher, since a smaller η_T value compels AURORA to keep ON a larger number of underutilized SCs. For instance, at $\eta_T = 0.6$, AURORA switches ON the next SC as soon as the utilization of the current ON SCs reach 60%. Thus, on average, a large number of SCs will be turned ON for smaller η_T values, thereby increasing energy consumption. Moreover, with a large number of SCs turned ON, there is a higher chance

that location estimation inaccuracy will result in turning ON SCs with low or no loads (i.e., high ECR [Joules/bit]). On the other hand, larger values of η_T enable AURORA to switch OFF a large number of SCs. For instance, at $\eta_T = 1$, AURORA will switch ON the next SC only when the utilization of the current ON SCs reaches 100%. As a result, the ECR is expected to decrease, and the same trend is observed for NARN. It is interesting to observe that, on the one hand, with an increasing value of η_T, a fewer number of SCs are turned ON; therefore, there is less chance of having any turned ON SCs with low or no load. On the other hand, with increasing η_T values, AURORA switches ON the smallest possible number of SCs, and all of them are almost fully utilized with few resources to spare. As a result inaccuracy in location estimation will result in an increased risk of blocking of the UEs (and hence an increased number of unsatisfied users–see Fig. 5.7), thereby negatively affecting the QoS. However, the number of fully utilized SCs is a more dominant factor in determining the overall ECR, as compared to a slight increase in the number of unsatisfied users; therefore, the overall ECR reduces. The comparison of AURORA with the reactive scheme demonstrates that the ECR for the reactive scheme is higher compared to AURORA. This is because in the reactive scheme, due to delayed user location information outdated configuration settings that are suboptimal for the current instant are applied to the network. This increases the percentage of unsatisfied users (on average, 1.85% with AURORA at $\eta_T = 1$, and 4% with the reactive scheme at high traffic loads), and the ECR is hence higher. Moreover, the ECR for AllOn-HomNet is slightly higher, compared to AllOn-HetNet. This is because higher CIO values used in AllOn-HetNet compel SCs to be utilized more, hence resulting in a reduced ECR, compared to an AllOn-HomNet scheme.

Fig. 5.5 illustrates the average number of SCs put to sleep mode with AU-

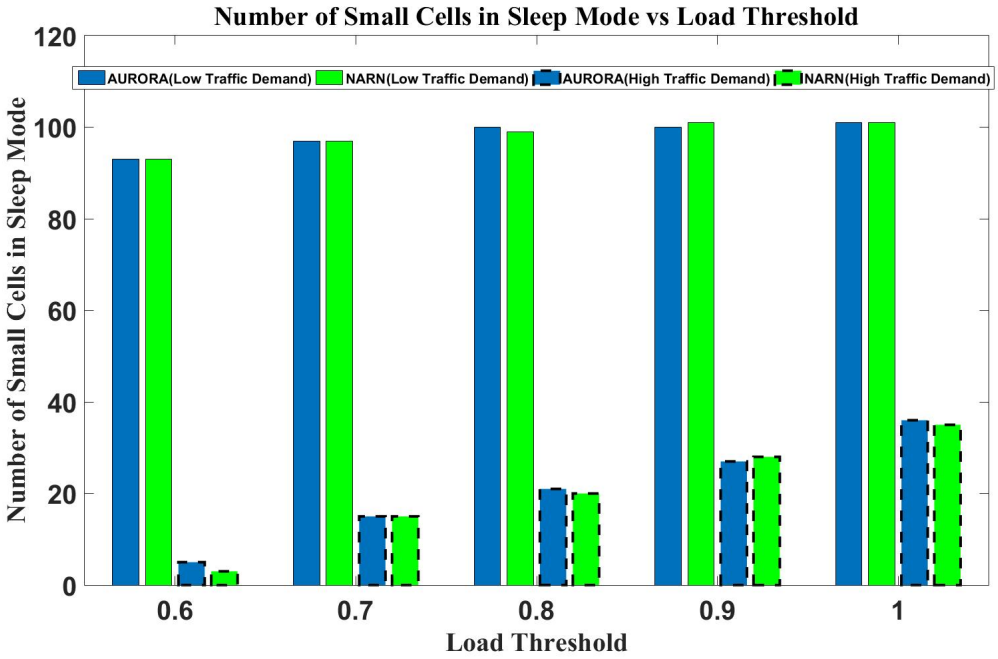
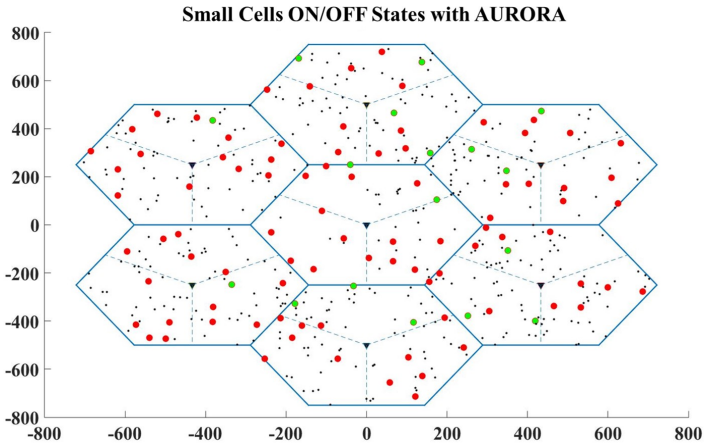
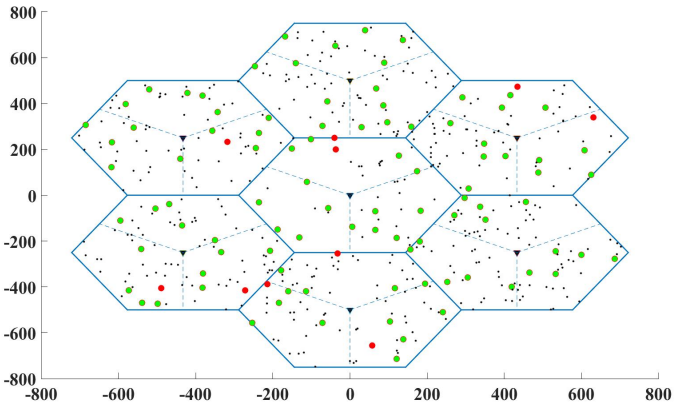


Fig. 5.5: Number of SCs put into sleep mode vs. load threshold

RORA and NARN with varying values of η_T for low and high traffic demand. It can be seen that a fewer number of SCs can be put into sleep mode to meet the needs of high traffic demand. The number of SCs put into sleep mode continue to increase with η_T. This is because with increasing values of η_T, an SC is utilized more before turning ON the next SC; i.e., more SCs are put into sleep mode at higher values of η_T. Since load coupled interference also increases with η_T, the optimization algorithm returns such an optimization parameter configuration (OPC), i.e., $\boldsymbol{\pi^c}, \boldsymbol{P^c_{CIO}}$, that minimizes the overall energy consumption ratio. Figure 5.6 presents a snapshot of the SCs states with AURORA for low and high traffic scenarios at the same time instants. It can be observed that for high traffic demand, the majority of the SCs are turned ON. Without a loss of generality, the results in all subsequent figures of this chapter correspond to only a high traffic demand scenario, which follows the same trend as that observed with low traffic demand. The average percentage of satisfied users under the AURORA framework vs load threshold η_T for a



(a) Low traffic demand



(b) High traffic demand

Fig. 5.6: Snapshot of small cell (ON/OFF) states by AURORA for (a) Low traffic demand and (b) High traffic demand. Green (red) circles indicate ON(OFF) SCs and UEs are illustrated by black dots.

high traffic demand scenario is visualized in Fig. 5.7 on the left y-axis, while EE ($1/ECR$) is plotted on the right y-axis. It can be observed that at low η_T values, plenty of free resources are available in a relatively higher number of available BSs, and more users are hence served with enough resources to meet their minimum QoS requirements. Even with location estimation inaccuracies, the UEs will still have a better chance of both acquiring enough resources and being satisfied. However, more SCs are turned ON at a low η_T with a higher chance of being underutilized, thereby resulting in a lower EE.

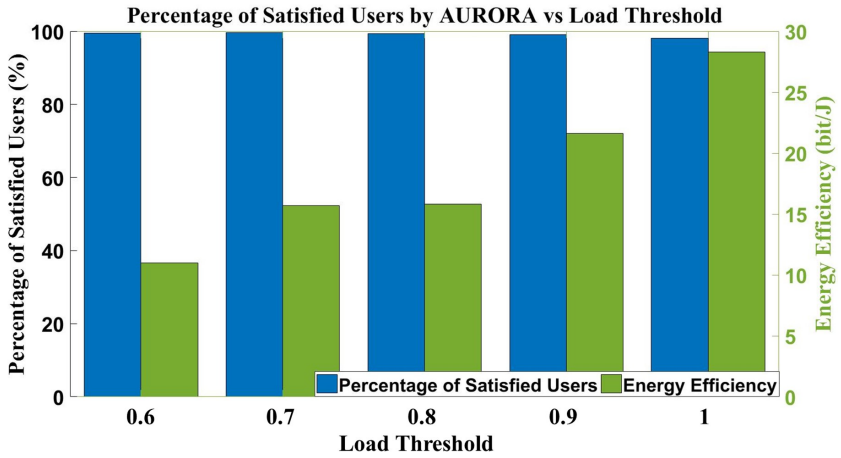


Fig. 5.7: Percentage of satisfied users vs. load threshold for high traffic demand

As the η_T value becomes higher and approaches 1, AURORA returns an OPC $\boldsymbol{\pi^c, P^c_{CIO}}$ that results in the smallest possible number of switched ON SCs, and all of them are almost fully utilized, with few resources to spare. Therefore, a slight location estimation inaccuracy can result in an increased risk of blocking and hence a decrease in the number of satisfied users. In contrast to that, fewer cells turned ON with more utilization improve the EE of the network. It is interesting to observe that for a high traffic demand scenario, even at $\eta_T = 1$, the percentage of satisfied users is above 98%. Figure 5.8 plots the cell loads of ON cells achievable with AURORA and NARN with $\eta_T = 0.6$ and 1 alongside AllOn schemes for high traffic demand.

It is evident from the figure that in the case of AllOn-HomNet and AllOn-HetNet, since all cells are kept ON, most of the cells are underutilized, with mean utilizations of 7.74% and 8% respectively. This results in a higher ECR (see Fig. 5.4). With AURORA and NARN, at a lower value of η_T, i.e., 0.6, some SCs are switched OFF, and the utilization of the remaining ON cells relatively increases with mean utilization of 30.9% and 27.6% respectively. At a higher value of η_T, i.e., 1, large number of SCs are switched OFF, and the few that are ON, are relatively more utilized, with mean utilizations of

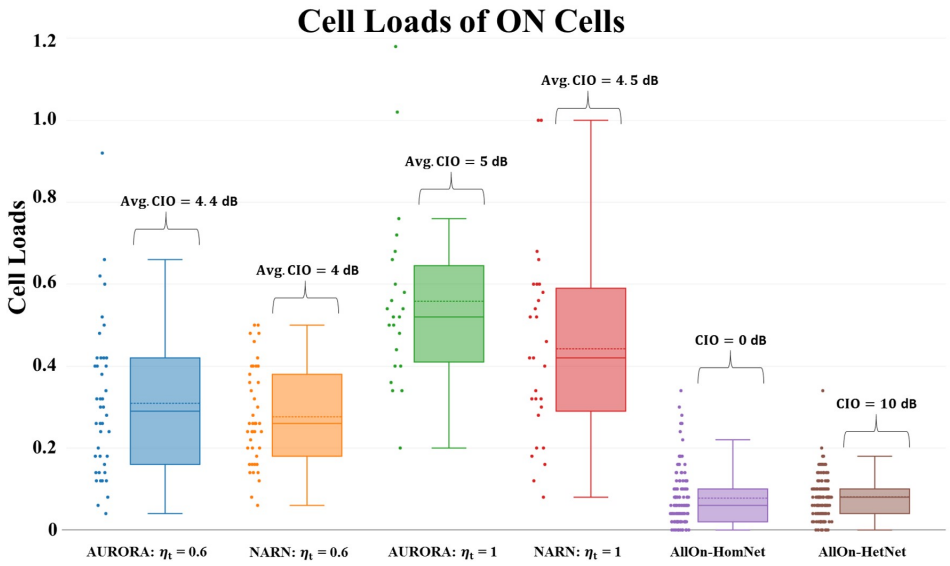


Fig. 5.8: Cell loads of ON cells for high traffic demand

55.8% and 44.2% respectively. The average CIO values are indicated on top of each boxplot. It is observed that a at higher η_T value of 1, as compared to a lower value of 0.6, on average, relatively larger CIO values have been leveraged. This is because when fewer cells are switched ON, the CIO values of ON SCs are boosted to serve the users of OFF cells. In this way, CIOs complement the proactive energy consumption optimization by serving as a guiding parameter in directing users to suitable cells such that the overall ECR reduces while satisfying QoS requirements. Fig. 5.9 presents a CDF plot of the results for the average downlink SINR for AURORA and NARN with $\eta_T = 0.6$ and 1, along with the AllOn-HomNet and AllOn-HetNet for a high traffic demand scenario. It can be observed that at a higher value of η_T, i.e., 1, load-coupled interference from neighboring BSs is high. Therefore, SINR is negatively affected for AURORA and NARN, as compared to AllOn-HomNet and AllOn-HetNet. As a matter of fact, when CIOs are leveraged, a degraded SINR is natural outcome. However, this does not mean a degraded system-wide performance as long as the loss in throughput caused by a lower logarithmic SINR term is offset by an increased number of PRBs allocable

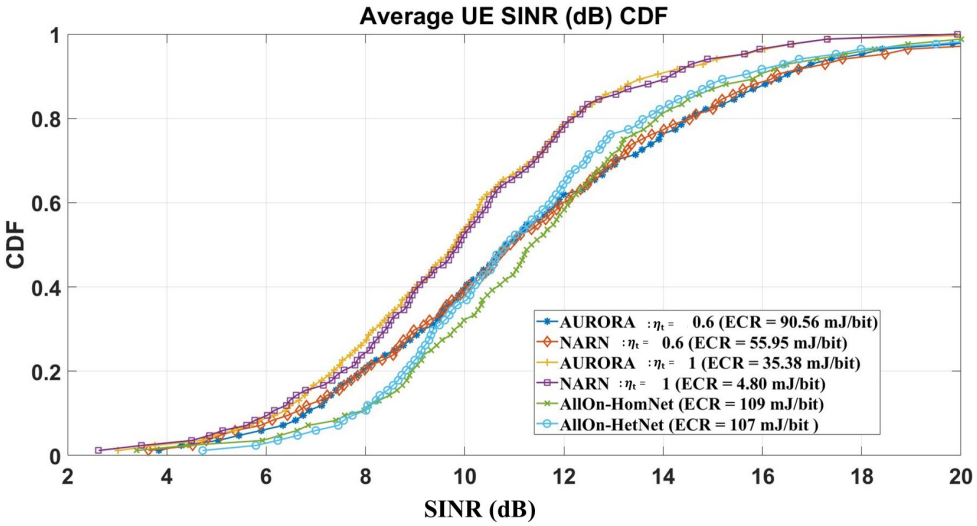


Fig. 5.9: Average UE SINR CDF for high traffic demand

to users. This is how AURORA strives to guarantees the minimum QoS requirements, as illustrated in Fig. 5.7. At a smaller η_T value of 0.6, a higher number of SCs are turned ON with a relatively less load. This reduces the overall interference floor in the network and the SINR improvement is hence higher than that achievable at an η_T value of 1. For AllOn-HomNet and AllOn-HetNet schemes, all SCs are ON and highly underutilized, resulting in a higher SINR. However, it is worth noting that this gain in SINR comes at the cost of a higher energy consumption; i.e., for AllOn-HomNet and AllOn-HetNet, the ECR is 109 mJ/bit and 107 mJ/bit respectively; this is much higher compared to the ECR for AURORA, which is around 36 mJ/bit, achievable at $\eta_T = 1$.

The average long-term cell occupancy probability of the users computed through (4.17) is depicted in Fig. 5.10(a) according to which users spend most of their time in macro cells 5, 1, 19, 20, and 21 (denoted by yellow stars). This information can be utilized for validation of the proposed AURORA framework. The average percentage of ON SCs with AURORA for a 1-hour simulation duration is presented in Fig. 5.10(b). As is evident, a higher number of SCs were turned ON in macro cells 9, 20, 5, 19, and 1 (denoted by yellow stars).

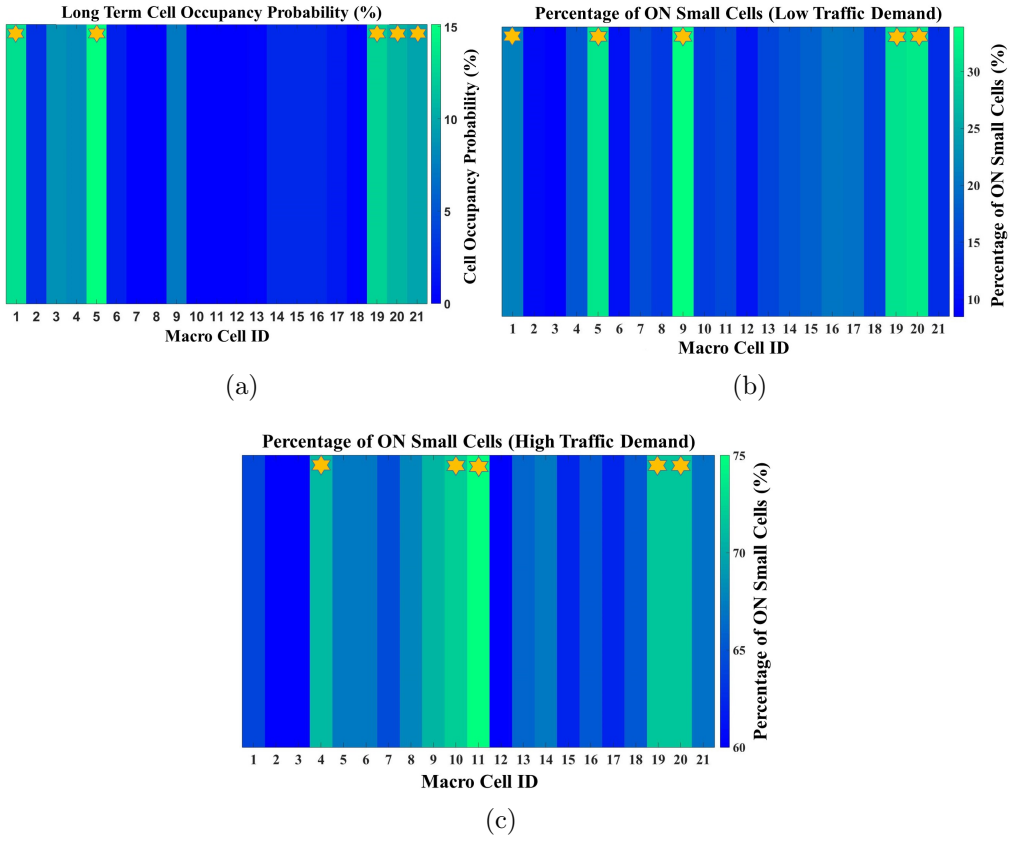


Fig. 5.10: (a) Long term cell occupancy probability (b) Percentage of ON small cells at low traffic demand (c) Percentage of ON small cells at high traffic demand.

Therefore, on average, AURORA kept a higher number of SCs switched ON in cells where users had a higher sojourn time. The few discrepancies that were observed, such as with macrocell 21, can be attributed to location estimation inaccuracies as well as the rate requirement of UEs in those cells; i.e., even with a higher cell occupancy probability of users in a particular macrocell, if the cumulative rate requirement of UEs is low, then SCs in that macrocell will remain switch OFF most of the time. For a higher traffic demand scenario, the average percentage of ON SCs with AURORA is depicted in Fig. 5.10(c). Since a higher number of SCs were turned ON to cope with the high traffic demand, the plot in Fig. 5.10(c) is relatively more green compared to that in Fig. 5.10(b).

5.3.3 Quantifying the Effect of Mobility Prediction Model Inaccuracy on Potential Energy Saving

The potential energy savings resulting from the the application of AURORA framework can be quantified by computing the ERG [96, 97] PM, given as follows:

$$ERG = (\frac{ECR_{Benchmark} - ECR_{AURORA}}{ECR_{Benchmark}}) \times 100\% \tag{5.11}$$

It is logical to anticipate that the ES gain of AURORA, i.e., the ERG, will depend on the accuracy of the underlying mobility prediction model. In this section, we analyze this dependence by varying the underlying user mobility model such that it includes varying degrees of randomness and hence predictability. To vary the degree of randomness in the mobility traces, the two key parameters of the SLAW mobility model, namely the variance in pause times and the percentage of random waypoints, were changed from the default values suggested in [69] (and used for the results in Figs. 5.4-5.10) to larger values to increase randomness in the mobility trajectory of the UEs. Four sets of gradually increasing initialization parameters were used that resulted in increasing randomness in user mobility. Our prediction model trained on these four sets of traces exhibited average prediction accuracies of 85%, 75%, 65% and 55%. The average ERG of AURORA for these varying values of prediction accuracy against AllOn-HomNet and AllOn-HetNet schemes, averaged over a 1-hour duration, for a high traffic demand scenario is plotted in Fig. 5.11. It is observed that, as expected, the gain of AURORA decreases with a decrease in prediction accuracy. However, it is noteworthy that as long as mobility is predictable with 55% or higher accuracy, AURORA continues to yield ERG. Given that typical human mobility features 93% predictability when averaged over a large real user sample space [53], AURORA is a promis-

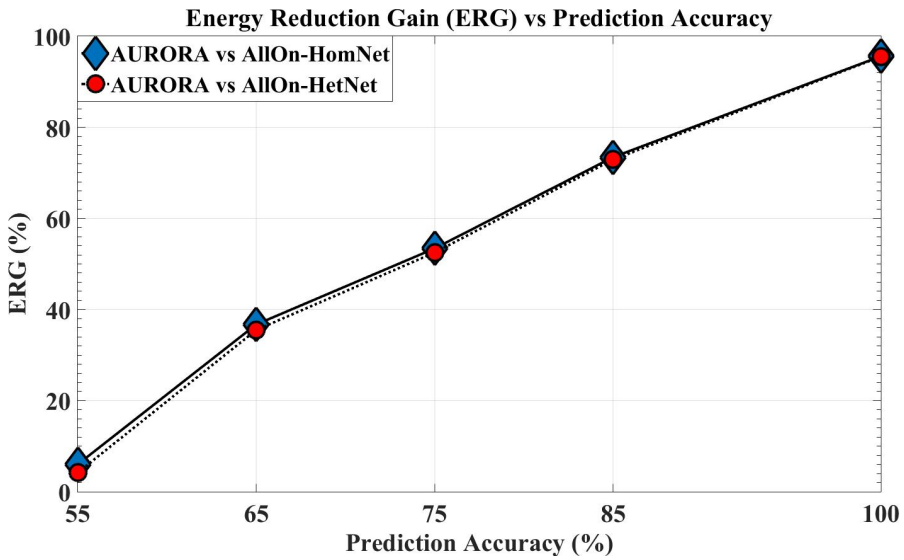


Fig. 5.11: Energy reduction gain vs prediction accuracy

ing approach. However, human mobility is bound to have some randomness that translates into prediction inaccuracy. The high frequency periodic update aspect of future location probabilities is one of the possible ways in which to cope with those prediction inaccuracies, as the effect of a prediction inaccuracy is only limited to the prediction interval. Another method is to make it adaptive so that AURORA continuously analyzes its performance and falls back on the conventional AllOn scheme when prediction accuracy drops below 55%. Moreover, selecting the top-two probable locations, as illustrated in Fig. 4.5, can also be chosen as a strategy to improve prediction accuracy, albeit at the cost of reduced ERG. Another approach is to use state-of-the-art machine learning algorithms in place of semi-Markov model. The AURORA framework is designed with flexibility in mind so that it can leverage any other machine learning technique for mobility prediction. To demonstrate this, we gauged the performance of the AURORA framework using deep neural network (DNN) as a mobility prediction model. The user's time stamped trajectory information was used as training data and mobility prediction was transformed into the classification problem with target cells as class labels.

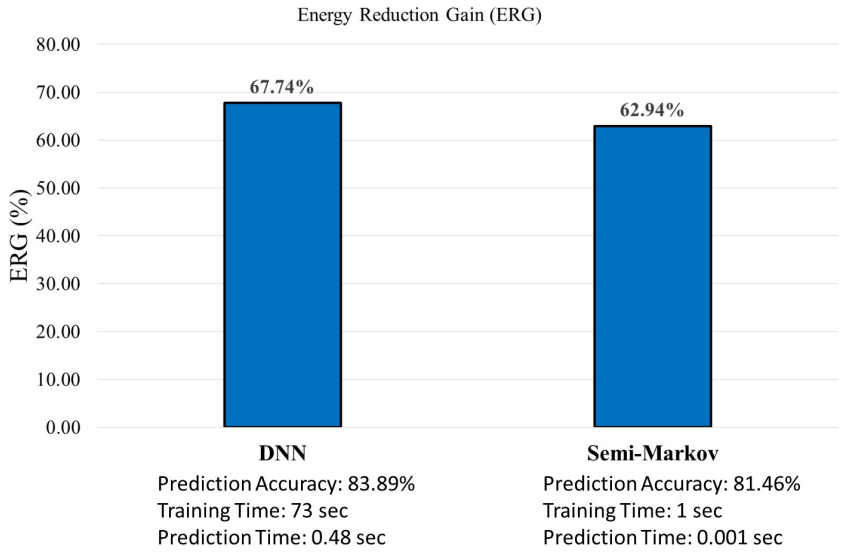


Fig. 5.12: Energy reduction gains with DNN and semi-Markov as mobility prediction models

Figure 5.12 illustrates the improvement in ERG (compared to AllOn-HetNet) when using alternative mobility prediction technique. It is observed that the gain of the proactive energy scheme increased with DNN albeit at the cost of increased time complexity.

5.4 Conclusion

This chapter proposed a novel spatiotemporal mobility prediction aware, proactive, sleep-mode-based ES optimization algorithm for cracking the future 5G ultra-dense HetNets puzzle. The proposed AURORA framework employs an innovative concept of estimating future user locations and leveraging them to estimate future cell loads. It then devises an ES optimization problem for the estimated future network scenario. The majority of conventional reactive-style approaches are expected to solve the formulated ES problem dynamically in real-time as network conditions change. However, this is close to impossible, even when substantial computing power is available. In contrast, the innovative proposed approach enables state-of-the-art heuristic techniques such

as GAs to find practically sound solutions to the formulated optimization problem predictively ahead of time. This advantage makes the proposed solution an enabler for meeting ambitious 5G latency and QoS requirements. Moreover, the AURORA framework considers the interplay between the three intertwined SON functions (ES, CCO, and LB) due to the overlap in their primary optimization parameters. Therefore, it employs a co-design approach wherein the joint optimization of ON/OFF states and CIO values for SCs does not conflict with CCO and LB objectives. Extensive simulations employing a realistic SLAW mobility model indicate that AURORA can achieve significant ERG in ultra-dense HetNets compared to the always ON approach. To test the sensitivity of AURORA to mobility prediction model and its accuracy, we investigate a DNN-based mobility prediction model as well. DNN-based mobility prediction model offers slightly higher prediction accuracy and hence better performance gain in AURORA, compared to semi-Markov but at the cost of substantial increased complexity and training time. A comparative performance analysis with a near-optimal performance bound indicates satisfactory robustness of the proposed AURORA framework for location estimation accuracies.

CHAPTER 6

Mobility Prediction-based, Proactive, Dynamic Network Orchestration for Load Balancing with QoS Constraint (OPERA)

> No one can whistle a symphony. It takes an orchestra to play it.
>
> Halford E. Luccock

Load imbalance among small and macro cells and consequential poor resource utilization is a major challenge that undermines the gains of emerging ultra-dense heterogeneous networks. While Load Balancing (LB) problem has been studied extensively, existing LB schemes in literature have one common caveat. They operate in reactive mode. i.e., cell parameters are tweaked reactively in response to changing cell loads. The inherent reactiveness of these LB schemes limits their ability to meet the extremely low latency and high QoE expected from future cellular networks vis-à-vis 5G and beyond. To address this challenge, in this chapter we propose a novel user mobility prediction based LB and network capacity optimization framework "OPERA". The proactiveness of OPERA stems from its novel capability that instead of passively waiting for congestion indicators to be observed and then reacting to them, OPERA predicts future cell loads using readily available data streams such as past HO traces, and then proactively optimizes key network parameters that affect cell load and network capacity namely azimuths, beam widths, Tx power and CIOs to preempt congestion before it happens. Although the resulted problem is NP-hard, the ahead of time estimation of cell loads allows

ample time for a dexterous combination heuristics such as genetic programming and pattern search to find solutions with high gain. We use extensive system level simulations to evaluate OPERA and compare its performance against three different benchmarks: (i) real network deployments settings taken from an LTE operator, (ii) recently proposed LB scheme in literature as representative of state-of-the-art reactive schemes, and (iii) upper performance bound where user future location is assumed to be known with 100% accuracy. Realistic SLAW model based mobility traces are used in the performance analysis. Results show that compared to benchmarks, OPERA can yield significant gain in terms of fairness in load distribution and percentage of satisfied users. Superior performance of OPERA on several fronts compared to current schemes stems from its following features: 1) It preempts congestion instead of reacting to it; 2) it actuates more parameters than any current LB schemes thereby increasing system level capacity instead of just shifting it among cells; 3) while performing LB, OPERA simultaneously maximizes residual capacity while incorporating throughput and coverage constraints; 4) it incorporates a load aware association strategy for ensuring conflict free operation of LB and CCO SON functions.

6.1 Introduction

The race to 5G is on with massive impromptu densification by small cells orchestrated by SON being perceived as a cost-effective solution to the impending mobile capacity crunch. Although poor indoor coverage coupled with explosive cellular data growth—that were expected to generate the momentous demand—are still relevant, to date, mass deployments of SCs remain elusive. One of the key challenge therein is the load imbalance issue [102]. Even with a targeted deployment, where these SCs are placed in high-traffic

zones, most users will still receive the strongest downlink signal from the tower-mounted macro cell. As a result, macro cells remain overloaded, while lightly loaded SCs are not able to serve more users–even those who are present in their coverage. This load imbalance also affects the user's perceived rate, which is the product of the instantaneous rate and the fraction of resources assigned to users. In case of highly loaded cells, few resources are assigned to users, and users' perceived QoE thus drastically falls. Therefore, load imbalance becomes an issue that is of paramount importance in HetNets.

6.1.1 Relevant Work

Load imbalance can be mitigated by shifting the traffic from high loaded cells to less loaded neighbors as far as interference and the coverage situation allow. To exploit this approach, LB has recently been adopted as a key SON function by 3GPP and has been extensively studied in literature [103, 104, 105, 106, 107, 108, 109, 110]. However, to the best of our knowledge, existing LB approaches fall short of the mark for 5G requirements due to the following limitations:

1. **Reactive mode of operation.** The plethora of existing LB SON algorithms are designed to mitigate load imbalance after detecting network conditions that have already taken effect. For example, when load imbalance is detected in a network, a non-convex, NP-hard LB algorithm is usually solved to optimize hard or soft network parameters. This is an improvement on fixed parameter settings in real networks that achieve LB at the cost of QoS. However, given the acute dynamics in HetNets, by the time load imbalance is detected and a realistic non-convex, NP-hard LB algorithm is solved to produce a new network configuration that is optimal for the observed network conditions, the conditions may have already changed. Therefore, the newly

determined optimal parameter settings are likely to be suboptimal before they can be actuated. This problem can be exacerbated, particularly in 5G, where myriad services and a plethora of cell types mean that the dynamics of a cellular eco-system will be even more swift.

2. Limited set of optimization parameters. Antenna tilts, downlink transmission power, and CIOs are the three prime optimization parameters that have been largely used in literature as actuators for the LB function. However, with the evolution of smart antenna technology, a new set of optimization parameters has surfaced that is yet to be exploited. This includes beam widths (radiation patterns) that can be adapted on the fly by optimizing the phases of complex weight vectors—thanks to smart antenna technology. Similarly, the azimuth orientation of the antennas can be leveraged to effectively change the cell footprint in conjunction with the antenna tilts, as illustrated in Fig. 1.2. As per the Sobol-based variance sensitivity analysis method [2], the first-order sensitivity index values for these optimization parameters are plotted in Fig. 1.3. It is observed that the CIOs, horizontal beam width, and azimuth are found to have the largest impact on network performance (the QoS). This observation calls for a deviation from the legacy age-old paradigm of only optimizing tilts and Tx power to maximize system performance and keeping other control knobs untouched.

3. SON conflict prone design. One caveat with conventional LB solutions is that they are oblivious to the fact that multiple SON functions may be prone to hidden or undesired conflict when implemented together in a network [3]. Another SON use case that becomes highly relevant to the load imbalance in HetNets is CCO because of the overlap of its optimization parameter set with LB. When CCO attempts to improve coverage by increasing Tx power, this can force a large number of users to jump into its coverage,

thereby conflicting with the LB SON objective. The interplay between CCO and LB becomes complicated, considering that both CCO and LB resort to optimization of the same parameters. Unlike antenna parameters, CIO is a soft parameter, and it was later introduced for LB and traffic steering in HetNets. However, an adjustment of CIO by the LB algorithm may also cause conflict with CCO objectives, since a user who is offloaded due to increased CIO may face higher interference (assuming intra-frequency offloading) and lower received power from the destination cell, compared to the origin cell. This may result in a lower SINR and ultimately lower throughputs. As explicated in [3], such a conflict-prone LB solution design can actually degrade a network's performance instead of improving it.

4. Impractical assumptions. There exist line of works, such as [111, 112, 113, 114], that are more theoretical in nature aimed for LB or more precisely optimal cell association in HetNets while considering CCO in form of constraints and vice versa. While these works provide valuable theoretical insights often into the asymptotic behavior of the system, for tractability, the analytical models used in these theoretical studies often build on overly-simplified and unrealistic assumptions such as uniformly distributed UEs, a spatially independent distribution of BSs, omnidirectional single-antenna transmission and reception, fixed transmit powers, the same CIO for all cells in one tier, and full load scenarios. These assumptions help to make the analysis tractable and the optimization convex in nature, but render the end result less useful for practical implementation. Contrary to dense HetNet as the main motivation for an LB SON function, some works on LB exist, such as [115, 99], wherein the solution is proposed and simulated mainly for macro-cell scenarios; i.e., large CIOs and Tx power disparities between SCs and macro cells are not considered. These approaches may work for current macro cell dominated

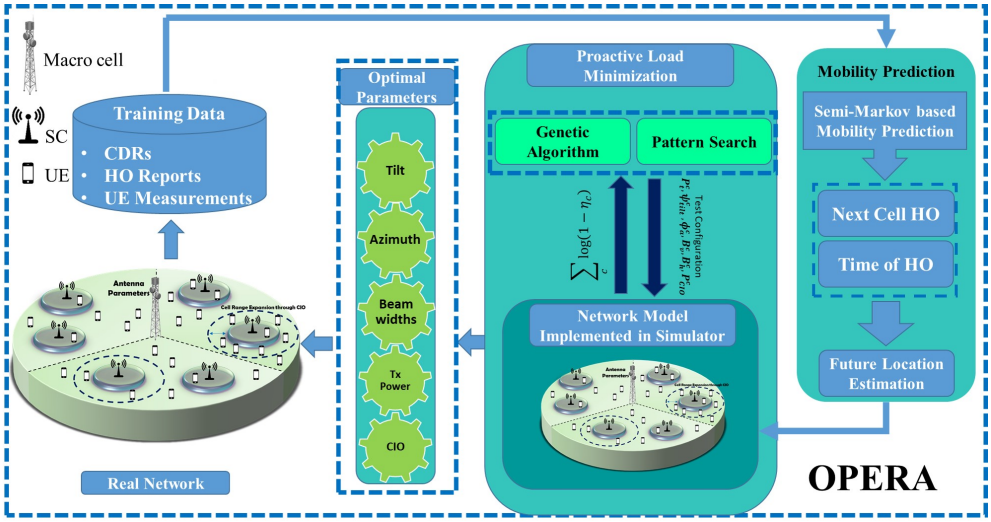


Fig. 6.1: OPERA framework

network deployment, but may not be applicable to dense HetNet envisioned for 5G.

6.2 The OPERA framework

To address the aforementioned limitations, we propose the OPERA framework (Fig. 6.1). OPERA can perform LB without conflicting with CCO. By building on the idea of big-data-empowered SON framework [5], OPERA leverages a novel approach to transform mobility from a challenge into an advantage. It proposes a solution that can leverage the knowledge gained from mobility/hand-off patterns for coping with the load imbalance challenge in 5G. The key idea is to make emerging cellular systems artificially intelligent and autonomous so that they can anticipate user mobility behavior. This intelligence is then used to formulate a novel LB optimization problem that proactively optimizes network parameters while satisfying QoS requirements.

In this section, we present the analytical model development of the OPERA framework (so called because it is a composition of a number of optimization parameters that are combined and actuated into a coordinated performance).

The OPERA framework consists of three components:

- Semi-Markov based spatiotemporal next cell prediction, as presented in chapter 4.
- The mapping of next cell prediction to future location estimation, presented in chapter 4.
- Proactive load-minimization optimization based on future location estimation.

6.2.1 Network Model and Assumptions

The OPERA framework proposed in this chapter only focuses on the downlink of cellular systems for the sake of conciseness. It is assumed that all mobile devices and SCs have omnidirectional antennas with a constant gain in all directions, while macro cells have smart directional antennas. A frequency reuse of one is considered, and the same band is utilized by the macro cell and the SCs. A full buffer traffic model is used for each user; i.e., there are always data available to be sent for a user with a CBR service. A centralized C-SON architecture is assumed wherein a centralized server in the core network performs a system-wide proactive load-minimization optimization. Moreover, user reported measurements that include the location-stamped information of past cell transitions, such as cell IDs, HO failure reports, RSRPs, and call detail records are assumed to be available to the C-SON server. These measurements are then utilized to build and train spatiotemporal mobility prediction models for users.

6.2.2 Proactive Load-Minimization Optimization

Given the next probable HO tuple and estimated future location $l^u_{k+k'}$ for all users, we devise a load-minimization optimization problem for next time step $k + k'$ in such a way that the network load is minimized while satisfying the coverage KPI and QoS requirement of each UE located at its estimated future location $l^u_{k+k'}$ as well as satisfying the maximum loading constraint for each BS. The added advantage of targeting load minimization is that many QoS-related KPIs are monotonic functions of the average cell loads, e.g., the throughput per resource experienced on average, the mean delay in the cell, or the average number of service requests that are present at any point in time. Due to monotonicity, reducing the average cell load simultaneously improves all quantiles of the spatial throughput distribution and similar measures, and the LB-based objective function can capture the goals of the CCO objective too. Moreover, load minimization or LB increases the probability of the availability of free resources in all the cells, and this becomes advantageous for HetNets. To emphasize this point, consider a scenario, for instance a macro cell and SC wherein the macro cell is bearing a load of 50%, while the SC is at a load of 90%. If a mobile user enters the SC coverage area and requires more throughput, he will be handed over to the macro cell, as the SC is already close to its maximum load utilization. On the other hand, a load minimization approach with minimum throughput guaranteed will try to minimize the load utilization of the two cells in the first place. As result of LB, if the load utilization of both cells is at 70%, and a new user comes under the SC umbrella with a high throughput requirement, then the SC will be able to accommodate this new user due to the availability of free resources.

The SINR $\hat{\gamma}^c_u$ at an estimated user location $l^u_{k+k'}$ at time step $k + k'$ when associated with a cell "c" is defined as follows:

$$\hat{\gamma}_u^c(k+k') = [\frac{P_t^c G_u 10^{-1.2\left(\lambda_v\left(\frac{\theta_u^c-\theta_{tilt}^c}{\varphi_v}\right)^2+\lambda_h\left(\frac{\phi_u^c-\phi_a^c}{\varphi_h}\right)^2\right)}\delta\alpha\left(d_u^c\right)^{-\beta}}{\kappa+\sum_{\forall i\in\mathbb{C}/c}P_t^i G_u 10^{-1.2\left(\lambda_v\left(\frac{\theta_u^i-\theta_{tilt}^i}{\varphi_v}\right)^2+\lambda_h\left(\frac{\phi_u^i-\phi_a^i}{\varphi_v}\right)^2\right)}\delta\alpha\left(d_u^i\right)^{-\beta}}]_{k+k'} \tag{6.1}$$

where P_t^c is the transmit power of cell c; G_u is the gain of user equipment; λ_v is the weight of the vertical beam pattern of the transmitter antenna; θ_u^c is the vertical angle of the user u in cell c with respect to horizon; θ_{tilt}^c is the tilt angle of the serving cell's antenna (at $\theta_{tilt}^c = 0^0$, BS antenna faces the horizon); φ_v is the vertical beam width of the transmitter antenna of cell c; λ_h is the weighting factor for the horizontal beam pattern; ϕ_u^c is the horizontal angle of user u in cell c with respect to absolute north; ϕ_a^c is the azimuth of the antenna of cell c ($\phi_a^c = 0^0$ corresponds to the absolute north); φ_h is the horizontal beam width of the transmitter antenna of cell c; δ is the shadowing observed by the signal; α is the path loss constant; d_u^c represents the distance of the estimated user location of "u," i.e., $l_{k+k'}^u$ from cell c; β is the path loss exponent; and κ is the noise variable. The time subscript on the right hand side of (6.1) and in the rest of the chapter indicates that all terms enclosed within $[.]_(k+k')$ are considered for the next time step $k+k^{'}$. Within scope of this chapter, it is assumed that shadowing estimate information for the estimated user location is available with a normally distributed error. In a practical network, channel maps that build on the MDT reports recently standardized by 3GPP and the collected channel quality indicator measurements can be utilized to estimate channel gains in estimated locations. This $\hat{\gamma}_u^c(k+k^{'})$ is a fully loaded SINR expression and is valid only when all cells are fully utilized. The actual interference from neighboring cells based on their respective loads is utilized as follows to calculate the SINR for data transmission:

$$\gamma_u^c(k+k') = [\frac{P_t^c G_u 10^{-1.2\left(\lambda_v\left(\frac{\theta_u^c-\theta_{tilt}^c}{\varphi_v}\right)^2+\lambda_h\left(\frac{\phi_u^c-\phi_a^c}{\varphi_h}\right)^2\right)}\delta\alpha\,(d_u^c)^{-\beta}}{\kappa+\sum_{\forall i\in\mathbb{C}/c}\eta_i P_t^i G_u 10^{-1.2\left(\lambda_v\left(\frac{\theta_u^i-\theta_{tilt}^i}{\varphi_v}\right)^2+\lambda_h\left(\frac{\phi_u^i-\phi_a^i}{\varphi_h}\right)^2\right)}\delta\alpha\,(d_u^i)^{-\beta}}]_{k+k'} \tag{6.2}$$

where η_i denotes the cell load in a cell i at time step $k+k'$ given by (5.5). This way of weighting the interference power received from each cell with its current resource utilization yields a certain coupling of the total interference with different cell utilizations. More loaded cells contribute more interference power than less loaded ones. The set of users connected to cell c is determined by the user association criterion defined in (5.6). Moreover, in addition to (5.6), in this work, we also leverage the user association criterion proposed by us in [116, 117] that takes the cell load into consideration; it is defined as follows:

$$\mathbb{U}_j := \left\{\forall u \in \mathbb{U} \mid j = \arg\max_{\forall c\in\mathbb{C}}\left(\left(\frac{1}{\eta_c}\right)^a * \left(P_{r,u_{dBm}}^c + P_{CIO_{dB}}^c\right)^{(1-a)}\right)\right\} \tag{6.3}$$

where η_c is the cell load, and $a \in [0,1]$ is the weighting factor to associate a level of priority to the load and RSRP metrics. A large value of a forces users to avoid highly loaded BSs, even if they provide good RSRP. Note that setting $a = 0$ will make it equivalent to (5.6). As such, a user is associated with a cell with which the product of the received power $(P_{r,u_{dBm}}^c + P_{CIO_{dB}}^c)$ and reciprocal of the cell load is maximum. Note that for the cell association criterion, η_c cannot be 0; therefore, for unloaded cells, η_c can be set as a very small number $\epsilon \to 0$. It is important to highlight here that in case of LB optimization with guaranteed minimum QoS requirements, it does not make sense to look at throughputs, since the UEs either obtain the exact CBR or are un-satisfied. A more appropriate PM to analyze hence is the number of

unsatisfied or dropped users "N_{us}," given in (5.7) [115].

Now, we formulate the general load minimization problem for time step $k+k'$ as follows:

$$\min_{P_t^c,\theta_{tilt}^c,\phi_a^c,\varphi_v^c,\varphi_h^c,P_{CIO}^c} \sum_c [-\log(1-\eta_c(P_t^c,\theta_{tilt}^c,\phi_a^c,\varphi_v^c,\varphi_h^c,P_{CIO}^c))]_{k+k'} \tag{6.4}$$

$$\min_{P_t^c,\theta_{tilt}^c,\phi_a^c,\varphi_v^c,\varphi_h^c,P_{CIO}^c} \sum_c -\log\left[1-\frac{1}{N_b^c}\sum_{u_c}\frac{\hat{\tau}_u}{\omega_B\log_2(1+\frac{P_t^c G_u 10^{-1.2\left(\lambda_v\left(\frac{\theta_u^c-\theta_{tilt}^c}{\varphi_v}\right)^2+\lambda_h\left(\frac{\phi_u^c-\phi_a^c}{\varphi_h}\right)^2\right)}\delta a(d_u^c)^{-\beta}}{\kappa+\sum_{\forall i\in\mathbb{C}/c}\eta_i P_t^i G_u 10^{-1.2\left(\lambda_v\left(\frac{\theta_u^i-\theta_{tilt}^i}{\varphi_v}\right)^2+\lambda_h\left(\frac{\phi_u^i-\phi_u^i}{\varphi_h}\right)^2\right)}\delta a(d_u^i)^{-\beta}})}\right]_{k+k'} \tag{6.5}$$

where

$$\mathbb{U}_j := \left\{\forall u\in\mathbb{U}\mid j=\arg\max_{\forall c\in\mathbb{C}}\left(\left(\frac{1}{\eta_c}\right)^a * \left(P_{r,u_{dBm}}^c + P_{CIO_{dB}}^c\right)^{(1-a)}\right)\right\} \tag{6.6}$$

subject to:

$$P_{t,min} \le P_t^c \le P_{t,max} \forall c\in\mathbb{C} \tag{6.7}$$

$$\theta_{min} \le \theta_{tilt}^c \le \theta_{max} \forall c\in\mathbb{C} \tag{6.8}$$

$$\phi_{min} \le \phi_t^a \le \phi_{max} \forall c\in\mathbb{C} \tag{6.9}$$

$$\varphi_{v,min} \le \varphi_v^c \le \varphi_{v,max} \forall c\in\mathbb{C} \tag{6.10}$$

$$\varphi_{h,min} \le \varphi_h^c \le \varphi_{h,max} \forall c\in\mathbb{C} \tag{6.11}$$

$$P_{CIO,min} \le P_{CIO}^c \le P_{CIO,max} \forall c\in\mathbb{C} \tag{6.12}$$

$$\frac{1}{|\mathbb{C}|}\sum_{\mathbb{C}}\frac{1}{|\mathbb{U}_c|}\sum_{\mathbb{U}_c} 1(P_{r,u}^c \ge P_{th}^c) \ge \bar{\omega} \tag{6.13}$$

$$\tau_u \ge \hat{\tau}_u \forall u\in\mathbb{U} \tag{6.14}$$

$$\eta_c < 1 \forall c\in\mathbb{C} \tag{6.15}$$

Since η_c denotes the resource utilization of cell "c", term $(1-\eta_c)$ represents the amount of resources available at cell "c" hence forth noted as residual capacity. The objective is to optimize the parameters P_t^c, θ_{tilt}^c ,ϕ_a^c ,φ_v^c,φ_h^c, and

P^c_{CIO} such that logarithmic sum of the idle resources in all cells is maximized while ensuring coverage reliability and the satisfaction of user throughput requirements. The log utility function leads to a kind of proportional fair treatment of the individual cells. The first six constraints define the limits for the variation in the Tx power, tilts, azimuths, beam widths (vertical, horizontal) and CIOs respectively. These are the constraints that will determine the size of the solution search space. The seventh constraint is to ensure minimum coverage, where P^c_{th} is the threshold for the minimum received power for a user to be considered covered, $\bar{\omega}$ defines the area coverage probability (a QoS KPI) that an operator wants to maintain, and 1(.) denotes the indicator function. The eighth constraint ensures that each user receives the required minimum bit rate depending on the QoS requirements of the service and the user's subscription level. This is due to the fact that to achieve the LB objective, the CIO of less loaded SCs may be increased to offload users of relatively more loaded cells into their coverage umbrella. The consequences are that the received power $P^c_{r,u}$ of offloaded users may become worse, leading to degraded SINR and throughputs. The effect of decreased SINR can be offset by allocating more resources only if the power received by the user is above a certain threshold. Therefore, this seventh constraint ensures that minimum throughput is guaranteed for all users in all cases, thereby inherently encompassing the CCO objective. However, this can only occur when the number of resources available in a cell is sufficient to meet user requirements; therefore, this constraint is complemented with a constraint on cell load $\eta_c < 1$.

The objective function, optimization variables, and constraints indicate that it is a large-scale, non-convex optimization problem due to the inherent coupling of optimization parameters and the cell loads. Non-convexity stems mainly from the fact that we are dealing with not one or two but six problem param-

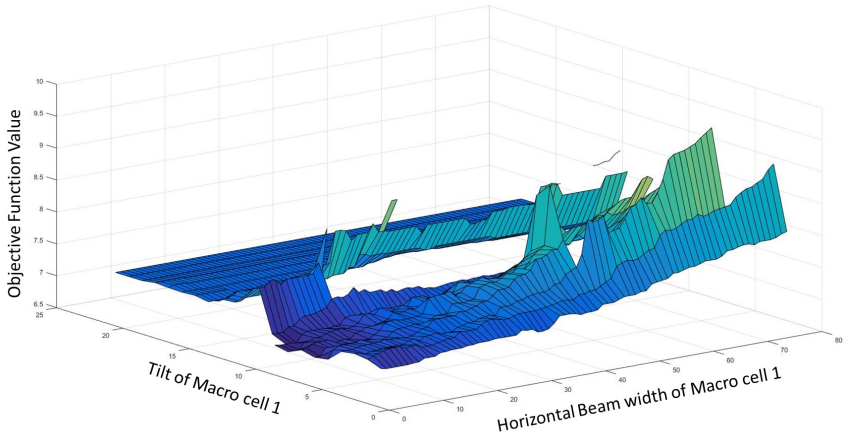


Fig. 6.2: Non-convexity behavior of the objective function

eters per cell, whose effects on the optimization function are not independent. The plot of the objective function for a sample topology of 42 cells is depicted in Fig. 6.2 wherein the tilt and horizontal beam width of a BS are varied, while the remaining variables are kept constant. It can be observed that the solution space is a combination of multiple hills and valleys (non-convex). As the number of possible combinations for the considered optimization parameters increases exponentially with network density, an exhaustive search for the optimal parameters to achieve the load minimization may not be practical for a large-sized network due to a high-complexity time search. For a practical scenario with 50 SCs and only CIO as an optimization variable with 10 possible values available at each SC, we already have 10^{50} possible settings. This is approximately equal to the number of atoms on earth. This search space size is too large to be traversed by a brute force algorithm in as short a time as a TTI. Therefore, to solve the formulated proactive LB problem for the next time step $(k + k')$ in real time, we utilized a hybrid combination of a GA and PS. As GA can reach the region near an optimum point relatively quickly, but it can take many function evaluations to achieve convergence; therefore, to overcome this issue, we used a hybrid scheme wherein the GA runs for a small

number of generations to reach a near optimum point. Then, the solution from the GA is used as an initial point for the PS algorithm that is faster and considered to be more efficient for local search. Pattern search methods proceed by conducting a series of exploratory moves about the current iterate before identifying a new iterate. Their pseudocodes [118] are given in the appendix.

Based on the estimated network state for time step $k + k'$, the OPERA framework consequently devises optimal values for all of the optimization parameters ahead of time such that LB is achieved. Optimization parameter values remain fixed from k to k'. As optimization algorithms need some time to converge; the proposed strategy allows ample time of k' duration to find a feasible solution.

6.3 Performance Evaluation

In this section, we present the results for our proposed OPERA framework. We have bench marked its performance against three schemes. (i) The first scheme comprises real mobile network deployment settings—RDS-A, RDS-B, and RDS-C are the three most common configurations adapted from real network LTE deployment settings for one of USA's national mobile operator in city of Tulsa with RDS-A (Tilt: 3^0) and RDS-B (Tilt: 5^0) both using antenna [119] and RDS-C (Tilt: 4^0) using antenna [120]. (ii) The second scheme is a Joint algorithm (referred to as Joint1 in [99]) that is quite relevant and has inspired the proposed work wherein LB is achieved via tilts with coverage constraints. It is used as a representative of state-of-art reactive schemes simulated by delaying user location information; i.e., the scheme is implemented for location information from the previous 1 minute. Note that due to the use of virtual loads in our system, the user association from [99]

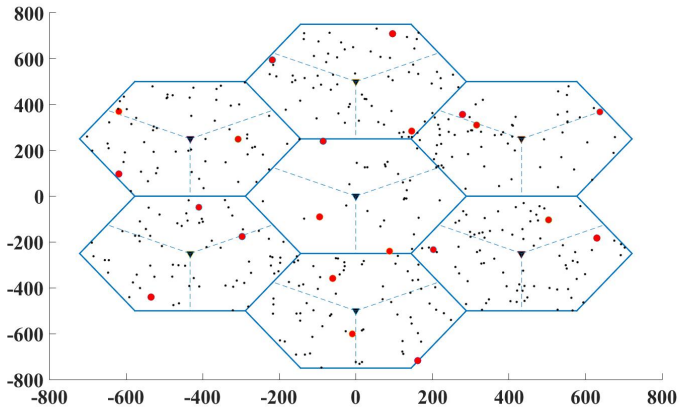


Fig. 6.3: Network topology with red circles indicating SCs, and UEs are illustrated by black dots

returns undefined results. Therefore, the algorithm in [99] is implemented using load-aware user association (6.3). (iii) The third scheme is near-optimal performance bound (NARN) wherein it is assumed that OPERA estimates a future location and the channel estimate at that location with 100% accuracy. NARN(OPERA) leverage a conventional association strategy ($a = 0$) while NARN*(OPERA*) uses a load-aware scheme with $a = 0.5$.

6.3.1 Simulation Settings

We generated typical macro cell and SC-based network and UE distributions leveraging an LTE 3GPP standard compliant [67] network topology simulator in MATLAB. The topology corresponding to one instant is illustrated in in Fig. 6.3, and the simulation parameter details are given in Table 6.1.

We used a wrap around model to simulate interference in an infinitely large network, thereby avoiding boundary effects. To model realistic networks, UEs were distributed non-uniformly in the coverage area such that a fraction of UEs were clustered around randomly located hotspots in each sector. Monte Carlo style simulation evaluations were used to estimate the average performance of

Table 6.1: Simulation parameter settings

System Parameters	Values
Number of Macro Base Stations	7 with 3 Sectors per Base Station
Small Cells per Sector	1
Number of UEs	Mobile: 84, Stationary: 336
Mobility Model	SLAW
Transmission frequency	2 GHz
Transmission Bandwidth	10 MHz
Network Topology	Hexagonal
Small Cell distribution	Uniform within Sector
UE distribution	Non- uniform with independent hotspots
UE Traffic classes	5 (Voice, Text Browsing, Image Browsing, FTP, Video) uniformly distributed
Macro Cell Tx Power	40 - 46 dBm
Macro Cell Tilt	90^0 - 120^0
Small Cell Tx Power	27 - 30 dBm
Small Cell CIO	Max: 10 dB, Min: 0 dB
Azimuths	-45^0 - 45^0
Horizontal Beam width	45^0 - 120^0
Vertical Beam width	5^0 - 15^0
Cellular System	LTE
Network Deployment Clutter	Urban
Macro Cell Height	25 m
Small Cell Height	10 m
UE Height	1.5 m
Inter-site Distance	500 m
Area Coverage Probability	100 %
Prediction Interval $k^{'}$	1 minute
Total Simulation Duration	60 minutes

the proposed framework. The SLAW model [69] was chosen as the mobility model, and it was utilized to generate HO traces of 84 mobile users for 1 week. Of this week, traces for the first six days were utilized to build and train the semi-Markov mobility model for each of the 84 UEs. Moreover, an additional 336 stationary UEs (80% of the total UEs [101]) were deployed to generate additional loading on the network. Without a loss of generality, we considered five different uniformly distributed UE traffic requirement profiles corresponding to desired throughputs of 24 kbps, 56 kbps, 128 kbps, 1,024 kbps, and 2,048 kbps. Without a loss of generality, and keeping operational complexity in mind, the prediction interval k' was set as 1 minute in our simulation study.

6.3.2 Results and Discussion

Figure 6.4 plots the histogram of difference (error) between predicted and actual load values with OPERA that leverage the semi-Markov-based future location algorithm presented in chapter 4. It is observed that most of the error falls into a 0.05 bin, with a root mean square error (RMSE) of 0.2711.

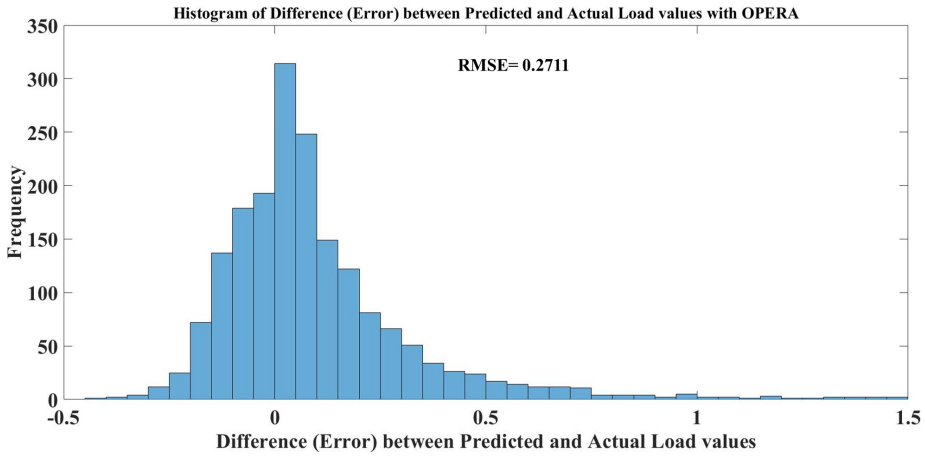


Fig. 6.4: Histogram of error between predicted and actual load values

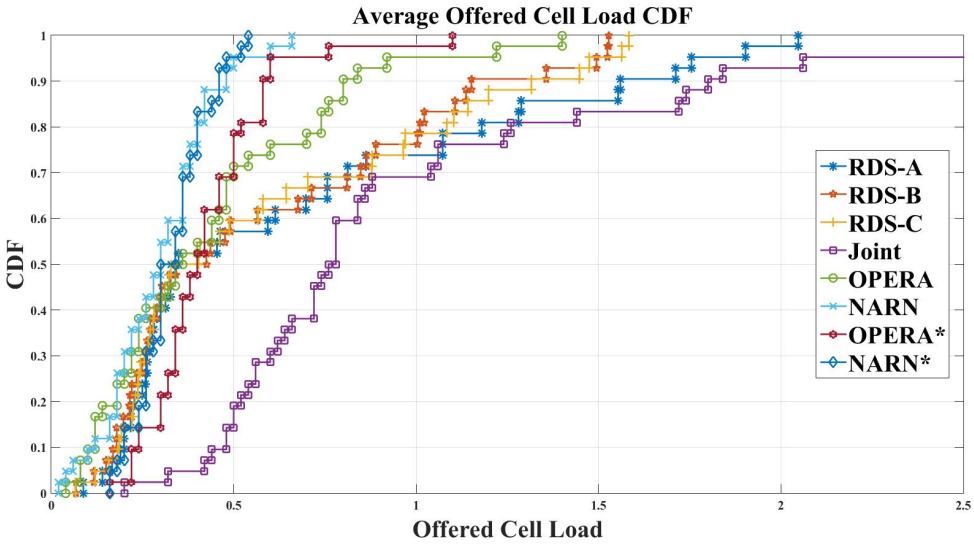


Fig. 6.5: Average offered cell loads CDF of all cells

Next, the offered cell load CDFs for all the cells with real deployment settings, the Joint scheme, and the proposed schemes are presented in Fig. 6.5. It is evident from the plot that with Joint scheme, the majority of the cells remain overloaded. The reason can be attributed to (i) a reactive approach, as resulting in outdated configuration settings, based on delayed user location information that is suboptimal for current instant, being applied to the network, and (ii) usage of only tilt as an optimization parameter. This increases the overloading or the percentage of unsatisfied users (as illustrated in Fig.

6.7). The same trend is observed for the real deployment settings wherein cells remain overloaded, with overloaded cells being maximum in RDS-A (around 26%), followed by RDS-C (around 23%), and RDS-B (around 21%). Compared to these fixed configuration settings and reactive schemes, the proposed solutions, namely, OPERA and OPERA*, achieve load reduction purely by increasing resource efficiency through the dexterous optimization of antenna parameters (transmission power, tilts, azimuths, and beam widths) and CIOs such that the cell loads are substantially reduced. However, slight overloading of around 4%(2%) is observed with OPERA (OPERA*) due to prediction inaccuracies. This overloading is mitigated when prediction accuracy reaches 100%, which is demonstrated by NARN and NARN* wherein the maximum cell loads observed are 66% and 54% respectively. It is observed that the inclusion of the load metric in the association criterion improves the residual capacity fairness in all cells, and as a result, even in the presence of prediction inaccuracies, cells have more free capacity to accommodate actual extra load as compared to a less predicted load. Figure 6.6 depicts the box plot of the percentage of free resources among all the cells, achievable with the RDS, reactive and proposed schemes. The inclusion of the load metric in the association criterion of OPERA* and NARN* demonstrates a somewhat tightly packed free resource values and hence less variance in residual capacity, compared to the rest of the schemes. Note that the OPERA and OPERA* show some cells with no free resources; this is due to prediction inaccuracies. This zero residual capacity scenario is avoided with NARN and NARN*. The variance in cells loads is further analyzed using Jain's fairness index, calculated through (6.16) and plotted in Fig. 6.7. In this figure, the average percentage of un-satisfied users is visualized on the left y-axis, while Jain's fairness index for residual capacity is plotted on the right y-axis, achievable with the RDS,

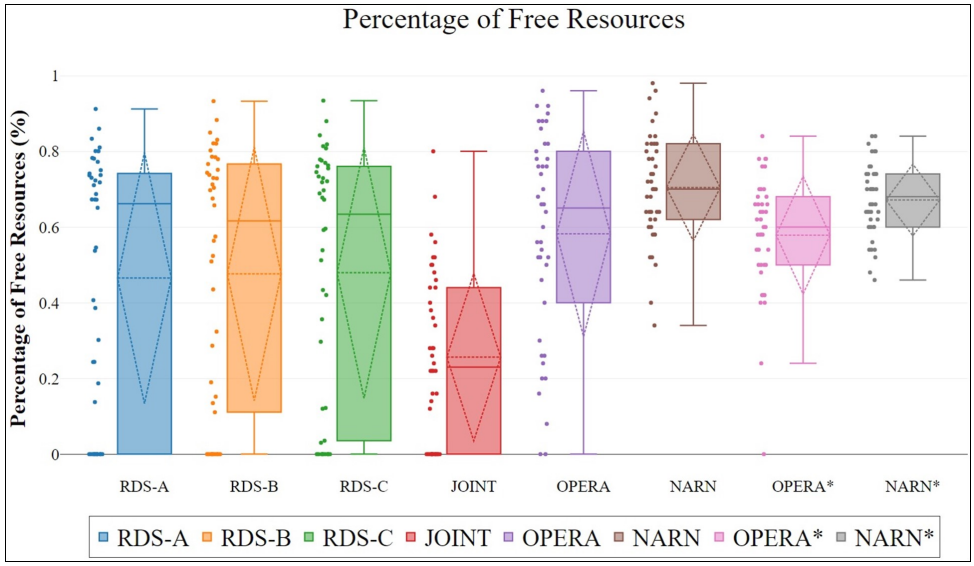


Fig. 6.6: Box plot of percentage of free resources in the cells

reactive and proposed schemes.

$$JFI(1-\eta_c) = \frac{(\sum_c(1-\eta_c))^2}{(|C| \times \sum_c(1-\eta_c)^2)} \tag{6.16}$$

The result computed from (6.16) ranges from (1/|C|) (worst case) to 1 (best case), and it is maximum when all the cells have the same amount of free residual capacity. Due to maximum overloading experienced with conventional RDS and reactive schemes, a considerable number of users face blocking and become unsatisfied. The proposed load-aware association-based schemes OPERA* (NARN*) achieve a maximum fairness of 0.967 (0.992), compared to their contemporaries OPERA (NARN), with a fairness of 0.965 (0.989). This fairness helps to reduce the percentage of unsatisfied users from 0.98% in OPERA to 0.35% in OPERA*. It is interesting to observe that even in the presence of imperfect prediction, the percentage of satisfied users is above 99% with OPERA.

Figure 6.8 plots the CDFs for the achievable UE SINRs with the RDS, reactive, and proposed schemes. For reactive and RDS schemes, SINR is con-

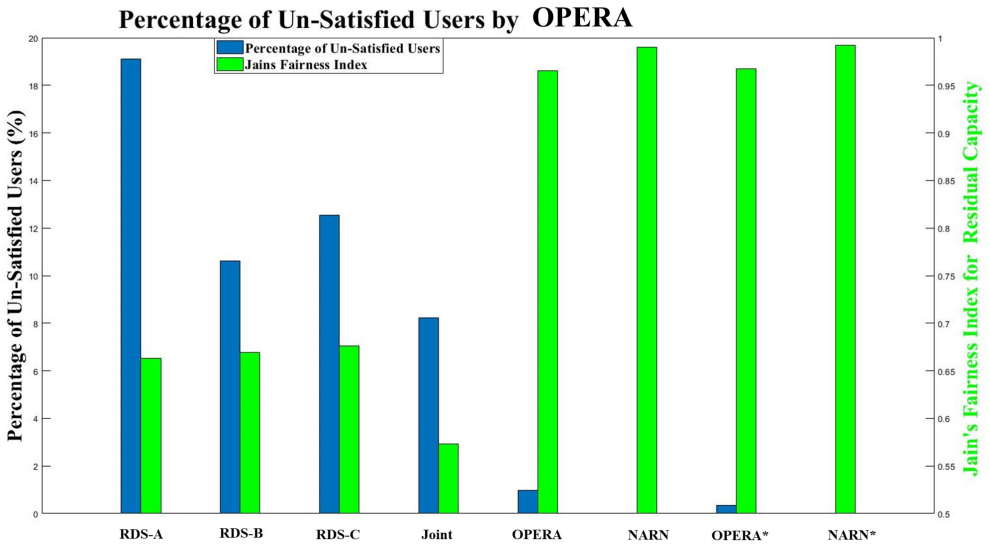


Fig. 6.7: QoE achieved with OPERA

siderably lower than the rest of the curves. The reason is that maximum loaded macro cells cause more network-wide interference, which reduces the achievable SINR of the UEs. This interference footprint of macro cells becomes highly contained with the proposed schemes (OPERA and OPERA*) by optimizing the values of antenna parameters and CIOs such that SINR is enhanced and cell loads are minimized. Moreover, the inclusion of the load metric in the association scheme (OPERA* and NARN*) reduces the achievable SINR of the UEs, as the UEs are not connected to the strongest possible cell. Despite decreasing SINR for NARN*, as compared to NARN, the solution manages to deliver the gains observed in Fig. 6.7, mainly because of load fairness by optimizing the horizontal and vertical beam widths, tilts, azimuths, Tx power, and CIOs. As a matter of fact, when CIOs are leveraged to increase system-wide capacity through LB, a degraded SINR is a natural outcome; however, it does not mean a degraded system-wide performance as long as the loss caused by the lower SINR is offset by an increased number of PRBs allocable to users. This compensating act is why OPERA* and NARN* outperform, hence the gain in resource utilization is observed.

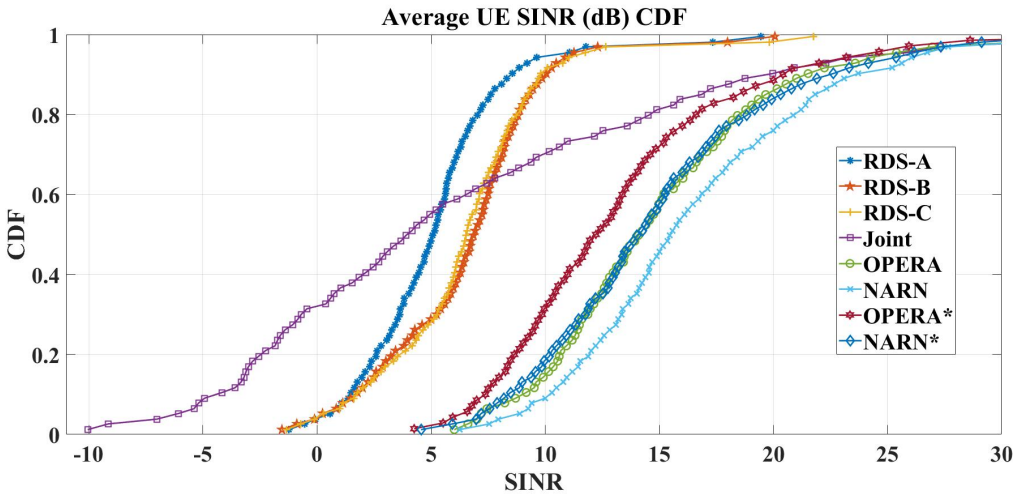


Fig. 6.8: Average UE SINR CDF

6.4 Conclusion

In this chapter, we proposed a novel spatiotemporal mobility prediction based proactive load balancing optimization framework for HetNets by jointly optimizing Tx power, tilts, azimuths, beam widths and CIOs. The proposed OPERA framework employs innovative concept of estimating future user locations and leverages that to estimate future cell loads. We then formulate a system-level fairness-aware load optimization problem for the estimated future cell specific loads. The majority of the current load balancing solutions are reactive and are expected to solve the LB problem dynamically in real-time after observing the congestion. With this reactive approach, meeting stringent QoS and latency requirements in 5G and beyond is close to impossible even when substantial computing power is available. Contrary to that, the innovative proposed approach enables state-of-the-art heuristic techniques like GA to find practically good solutions to the formulated optimization problem predictively ahead of time. Moreover, OPERA framework accounts for the interplay between two intertwined SON functions (LB and CCO) that have been shown to have strong conflict due to the overlap among their primary op-

timization parameters and thus ensures conflict free operation. A load aware association strategy that underpins OPERA further bolsters the framework against location estimation accuracies and maximizes system level capacity and QoE in addition to balancing load. Extensive simulations employing realistic SLAW mobility model indicate that, in best case, OPERA can reduce percentage of unsatisfied users to 0.35% making it an enabler for meeting 5G ambitious QoE requirements despite of acute mobility and heterogeneity of cell sizes. The presented results highlight the value of prediction (AI) based proactive optimization.

CHAPTER 7

Proactive Self-Healing for Future Cellular Networks

> The sciences do not try to explain, they hardly even try to interpret, they mainly make models. By a model is meant a mathematical construct which, with the addition of certain verbal interpretations, describes observed phenomena. The justification of such a mathematical construct is solely and precisely that it is expected to work.
>
> John von Neumann

This chapter presents a stochastic analytical model to analyze and predict the arrival of faults in a cellular network. It exploits CTMC with exponential distribution for failures and recovery times to model the reliability behavior of a BS. It then leverages the developed model and subsequent analysis to propose an adaptive fault predictive framework. The proposed fault prediction framework can adapt the CTMC model by dynamically learning from past database of failures, and hence can reduce network recovery time thereby improving its reliability.

7.1 Introduction

Cellular networks are inherently subject to cell outages caused by either BS hardware and/or software malfunctions or misconfiguration of several hundred cell parameters during routine network operation (Fig. 7.1). A BS can be susceptible to a complete outage due to critical failures, or it can exhibit degraded performance in case of trivial failures. Forthcoming cellular networks

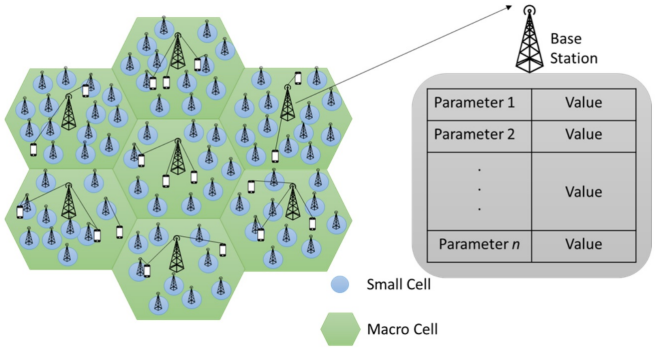


Fig. 7.1: Ultra-dense, heterogeneous, complex cellular network

are susceptible to even higher cell outage rates, as the multiple SON functions may be subjected to a large number of potential conflicts when operated concurrently in a system. Given the parametric overlap as well as coupling among the objectives of different SON functions, it has been demonstrated in [3] that a large number of conflicts are possible among SON functions if no self-coordination mechanism is employed. At times, such conflicts can actually degrade a network's performance instead of improving it. For example, the CCO SON function might try to improve coverage by increasing transmission power; however, this may conflict with the ES SON function. The potential failures occurring due to hardware/software malfunctioning, multi-vendor incompatibility, or SON conflicts ultimately affect the coverage and performance reliability of the network.

The increasing number of radio nodes in the 5G mobile cellular network can result in an increase in node failures [121]. This is demonstrated in Fig. 7.2, which illustrates the outage probability of a cell as mobile cellular network density increases, obtained using a Poisson distribution-based method for estimating node failures derived from [121]. Figure 7.2 displays the probability of a single node failure in one day (lower line chart), in three days (middle line chart), and in seven days (top line chart). We can see that the prob-

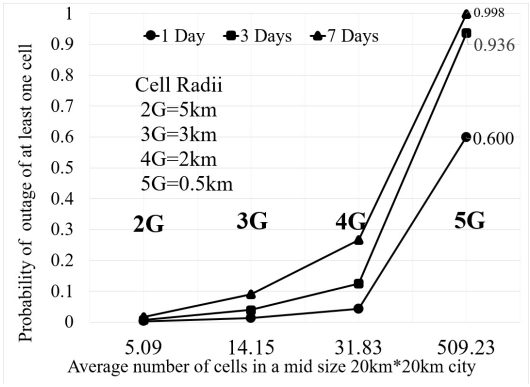


Fig. 7.2: Outage probability of one cell with increase in cell density

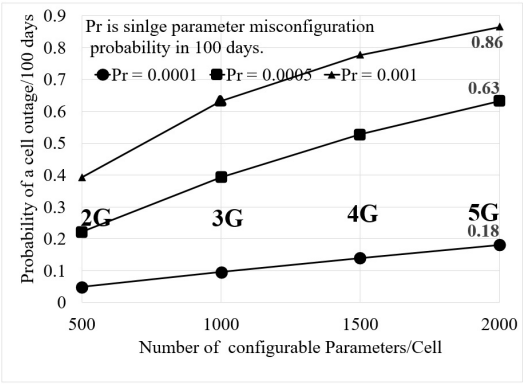


Fig. 7.3: Probability of single parameter misconfiguration with increase in number of configurable parameters

ability of node failures is relatively low in a low-density network such as a second-generation mobile cellular network. However, as the network density increases, the probability of node failure increases, so much so that on any given day, the probability of node failure could be anywhere between 60% and 99.8%.

The increasing number of network control parameters and entities can raise the probability of parameter misconfiguration significantly. A quantitative analysis of parameter misconfiguration in 5G mobile cellular networks is presented in Fig. 7.3, which illustrates the probability of misconfiguration of one parameter per cell every 100 days as the total number of configurable parameters per cell increases. The parameter misconfiguration probability is

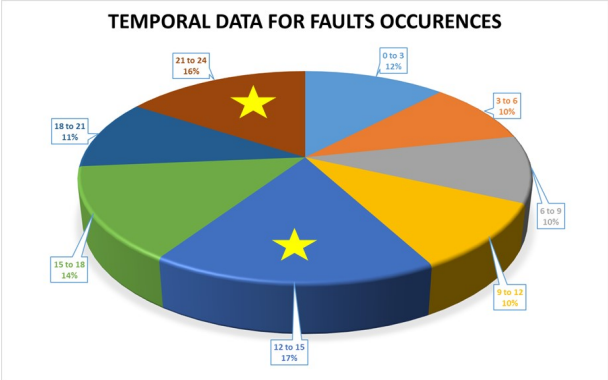


Fig. 7.4: Percentage of faults in given time interval

also derived using the Poisson distribution-based method of failure estimation presented in [121]. In Fig. 7.3, three different probabilities 0.01% (bottom line chart), 0.05% (middle line chart), and 0.1% (top line chart) of parametric misconfiguration per 100 days are assumed. From Fig. 7.3, it is clear that parametric misconfiguration will become a major concern for mobile network operators in 5G networks.

Until now, all SON self-healing systems have been passive or reactive systems that are only able to adapt to conditions many minutes after an event has taken place. This often means that changes are made to the network long after the need for change has passed, thereby creating a second negative impact to compound the first. This method only responds to problems and can only hope to limit their magnitude, not prevent them from happening. Furthermore, this method is not experiencing the best success, since the network experiences the greatest number of faults when people need to make calls the most. Figure 7.4 was generated from the data set, which was an array of time-stamped faults in a month from a national U.S. mobile operator. It illustrates that during lunch (usually time interval 12 to 15), the greatest number of faults occurred. Also, there were a significant number of faults at night (21 to 24) and after lunch (15 to 18). Regardless of the time interval, it is important

for customers to have a reliable network when they need to make calls.
This is why predictive analysis, which is a proactive approach, is necessary due to its effectiveness. Intending to deal with the problem before it occurs, predictive analysis techniques learn the behavior of the system under normal conditions, and they can also monitor the patterns that forecast a troublesome scenario. In telecommunications specifically, carriers would be able to estimate that a fault is going to occur in a certain amount of time, and they could thus take steps to prevent that fault. This would eliminate any possible difficulties that a customer would experience, and it is much better than the reactive approach, which can only hope to limit the trouble the customer must experience.

7.2 Reliability Analytical Model for Cellular Networks

The reliability analysis of future cellular network's BSs is of paramount importance for network operators, since it directly affects the QoS and user experience. A quantitative analysis of SON reliability can also provide vendors with better insight into the various reliability considerations in SONs. Furthermore, it can help to improve operators' confidence in SONs, which has been major bottleneck in SON penetration despite the significant financial and technical gains that SONs can offer.

7.2.1 Prior Works

Despite the great significance of the topic, few studies to date have focused on the reliability and survivability analysis of cellular networks in general and SON-enabled cellular networks in particular. Dharmaraja et al. [122] developed an analytical model for the reliability and survivability quantification

of a UMTS architecture network. Xie et al. [123] modeled and analyzed the survivability of an infrastructure based wireless network under disaster propagation. Tipper et al. [124] performed a simulation-based survivability analysis of a mobile network. Guida et al. [125] evaluated the performance of IP multimedia subsystem (IMS) core network signaling servers. However, unlike the previous works on cellular network reliability that mostly focus on the structural aspects of cellular networks and overlook the network behavioral aspects that can cause complete or partial failures, our work is more focused on developing a generic analytical model, encompassing diverse faults cases such as software/hardware failures or SON-conflict attributed misconfigurations. This approach allows for the flexibility to incorporate a variety of failure scenarios into the model. To the best of our knowledge, a study that analyzes the probabilistic reliability behavior of SON-enabled emerging cellular networks, including 5G, by considering the failure probability of BSs using CTMC does not exist in open literature. This contribution is thus a first attempt in that direction.

7.2.2 Model Development

To analyze and evaluate the reliability behavior of a cellular network, a quantitative model for a cell (a BS) failure is needed. In real-world cases, most of the node failure and repair times follow time-dependent failure rate distributions such as Weibull, Pareto, and lognormal [126]. However, in most cases, analytical models with general (non-exponential) distributions are not mathematically tractable. Therefore, a phase-type distribution, which is a convolution of many exponential phases, is used to approximate many general distributions and is used to construct the mathematically solvable analytical models [122], for a component reliability analysis [127]. Since exponential distribu-

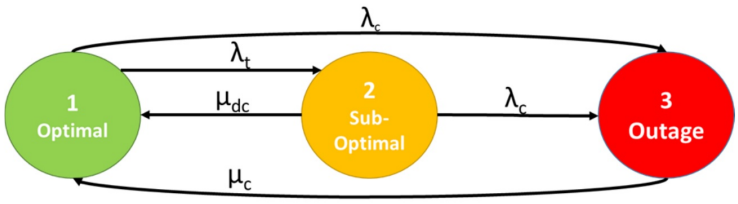


Fig. 7.5: State transition diagram for a SON-enabled BS

tion is a particular case of a phase type distribution, hardware and software faults are commonly modeled as an exponential distribution. Therefore, in this contribution, we consider that the time to transit from a system state to another due to failures and recovery also follows an exponential distribution. This assumption is also supported by the fact that the exponential random variable is the only continuous random variable with Markov property.

Building on this assumption, we construct an analytically tractable CTMC model for the reliability analysis of a SON-enabled BS. Figure 7.5 illustrates the state transition diagram of the CTMC model for the probabilistic reliability behavior of the BS.

Let $X(t)$, with finite state space $\mathfrak{S} = \{1, 2, 3\}$, denote the state of the BS at time t wherein $X(t) = 1$ if the BS is in a healthy state at time t, with all parameters configured with optimal values. $X(t) = 2$ if the BS is in a sub-optimal state at time t, with one or more of the parameters misconfigured. In this state, the cell continues to operate, but its performance degrades below a typical level of performance. $X(t) = 3$ if the BS is in complete outage at time t. It is assumed that time for failure is exponentially distributed. Since the rate of arrival of failures is temporarily independent, it can be modeled using Poisson distribution. We classify failures into (1) trivial failures characterized by arrival rate λ_t which are failures that do not cause complete outage but drive the network from an optimal to sub-optimal state, and (2) critical failures characterized by arrival rate λ_c which lead to complete outage of the

cell. Therefore, trivial failures can only occur when a network is in an optimal state, while critical failures can occur in state 1 or state 2. Each BS is assumed to be equipped with a recovery module powered by a self-coordination framework such as that proposed in [3]. This module reconfigures all configuration parameters back to their original optimal values once the misconfiguration is detected and diagnosed. Moreover, it has the capability to reset the BS software or switch over to the secondary backup hardware board if failure has stemmed from hardware/software-related issues. The time to move the network from a sub-optimal state back to an optimal state is assumed to be exponentially distributed with mean value $1/\mu_{dc}$. This includes the time for cell anomaly detection, diagnosis, and compensation [128, 129]. Similarly, the time period to recover from a complete outage is exponentially distributed with mean value $1/\mu_c$, which generally involves time for compensation only. Furthermore, the failure or repair transition is only determined by the current state and not on the path to the current state. With these assumptions, the transient process $X(t)$ can be mathematically modeled as a temporally homogeneous CTMC on the state space $\mathfrak{S}$. For each time $t > 0$, the probability that the BS is in state j is given by the following equation:

$$p_j(t) = Pr\{X(t) = j\}, j \in \mathfrak{S} \tag{7.1}$$

Let $p(t) = [p_1(t), p_2(t), p_3(t)]$ denote the row vector of the transient state probabilities of $X(t)$. The generator matrix $\mathbf{G}$ and rate matrix $\mathbf{V}$ for this CTMC $X(t)$ are given as follows:

$$\mathbf{G} = \begin{bmatrix} -\lambda_t - \lambda_c & \lambda_t & \lambda_c \\ \mu_{dc} & -\mu_{dc} - \lambda_c & \lambda_c \\ \mu_c & 0 & -\mu_c \end{bmatrix} \tag{7.2}$$

$$\mathbf{V} = \begin{bmatrix} 0 & \lambda_t & \lambda_c \\ \mu_{dc} & 0 & \lambda_c \\ \mu_c & 0 & 0 \end{bmatrix} \tag{7.3}$$

7.2.3 Analysis

In this section, we perform a transient analysis followed by the computation of PMs.

Transient Analysis

Using generator matrix $\mathbf{G}$, the dynamic behavior of the CTMC can be described by the Kolmogorov differential equation in the following matrix form:

$$\frac{d\mathbf{P}(t)}{dt} = \mathbf{P}(t)\mathbf{G} \tag{7.4}$$

Then, the transient state probability vector can be obtained as follows:

$$\mathbf{P}(t) = \mathbf{P}(0)e^{\mathbf{G}t} \tag{7.5}$$

where $\mathbf{P}(0)$ is the initial probability vector. The common methods to obtain the transient probability vector $\mathbf{P}(t)$ include the matrix exponential approach [130] and uniformization [131]. In this contribution, we resort to the uniformization method for the analysis because of its higher accuracy and efficient computation due to which it is the method of choice for typical problems similar to the one under consideration in this chapter [132]. Let g_{ii} be the diagonal element of $\mathbf{G}$, and let $\mathbf{I}$ be the unit matrix; then the transient state probability vector can be obtained as follows:

$$\mathbf{P}(t) = \sum_{k=0}^{\infty} e^{-\rho t} \frac{(\rho t)^k}{k!} \hat{\mathbf{P}^k} \tag{7.6}$$

where $\rho \geq \max_i |g_{ii}|$ is the uniform rate parameter, and $\hat{\mathbf{P}}$ is the probability transition matrix given as follows:

$$\hat{\mathbf{P}} = \mathbf{I} + \frac{\mathbf{G}}{\rho} \tag{7.7}$$

When we truncate the summation in (7.6) from infinity to some large number M, the resulting controllable accuracy error can be computed as follows:

$$\epsilon = 1 - \sum_{k=0}^{M} e^{-\rho t} \frac{(\rho t)^k}{k!} \tag{7.8}$$

Performance Metrics

Based on the uniformization method, three PMs are computed as follows to quantify the reliability of the network:

Occupancy Time:

The expected length of time that the BS spends in each of the states, namely, optimal, suboptimal, and outage, during a given interval of time can be determined using the occupancy time of the CTMC. Let $\Psi_{i,j}(T)$ be the expected amount of time that the CTMC spends in state j during the interval $[0, T]$, starting in state i, and let $p_{i,j}(t)$ be the element of the transition probability matrix $\hat{\mathbf{P}}$. The quantity $\Psi_{i,j}(T)$ is called the occupancy time of state j until time T, starting from state i given as follows:

$$\Psi_{i,j}(T) = \int_0^T p_{i,j}(t) dt \tag{7.9}$$

and in matrix form:

$$\Psi(T) = \begin{bmatrix} \Psi_{1,1} & \Psi_{1,2} & \Psi_{1,3} \\ \Psi_{2,1} & \Psi_{2,2} & \Psi_{2,3} \\ \Psi_{3,1} & \Psi_{3,2} & \Psi_{3,3} \end{bmatrix} \tag{7.10}$$

First Passage Time

The expected value of the random time at which a BS passes into each of the states (optimal, suboptimal, outage) for the first time can be calculated using the first passage times of the CTMC. The first-passage time ζ_j into state j starting from state i is defined as follows:

$$\zeta_j = E(T|X(0) = i) \tag{7.11}$$

where

$$T = \min\{t \geq 0 : X(t) = j\} \tag{7.12}$$

and E is the expected value. The first passage times for a CTMC with a state space $\mathfrak{S}$ satisfy the following relation [131]:

$$v_i\zeta_i = 1 + \sum_{j=1}^{N-1} v_{i,j}\zeta_j,\ 1 \leq i \leq N-1 \tag{7.13}$$

where $i, j \in \mathfrak{S}$ and $v_i = \sum_{j=1}^{N} v_{i,j}, \mathbf{V} = [v_{i,j}]$. Therefore, in our model, the first passage time for state 2 will be:

$$(\lambda_t + \lambda_c)\zeta_1 = 1 + \lambda_c\zeta_3 \tag{7.14}$$

$$(\mu_c)\zeta_3 = 1 + \mu_c\zeta_1 \tag{7.15}$$

By solving (7.14) and (7.15), we obtain the following equation:

$$\zeta_{3\to2} = \left(\frac{(\lambda_t + \lambda_c) + \mu_c}{((\lambda_t + \lambda_c) \times \mu_c) - \lambda_c\mu_c}\right) \tag{7.16}$$

$$\zeta_{1\to2} = \left(\frac{\mu_c\zeta_{3\to2} - 1}{\mu_c}\right) \tag{7.17}$$

where $\zeta_{3\to2}$ and $\zeta_{1\to2}$ are the first passage times to state 2 starting from states 3 and 1 respectively. The first passage time for state 3 will be

$$(\lambda_t + \lambda_c)\zeta_1 = 1 + \lambda_t\zeta_2 \tag{7.18}$$

$$(\mu_{dc} + \lambda_c)\zeta_2 = 1 + \mu_{dc}\zeta_1 \tag{7.19}$$

By solving (7.18) and (7.19), we obtain the following equation:

$$\zeta_{1\to3} = \frac{(\mu_{dc} + \lambda_c + \lambda_t)}{(\lambda_t + \lambda_c)(\mu_{dc} + \lambda_c) - \lambda_t\mu_{dc}} \tag{7.20}$$

$$\zeta_{2\to3} = \frac{1 + \mu_{dc}\zeta_{1\to3}}{\mu_{dc} + \lambda_c} \tag{7.21}$$

where $\zeta_{1\to3}$ and $\zeta_{2\to3}$ are the first passage times to state 3 starting from states 1 and 2 respectively.

Steady-State Distribution

To analyze the long-term behavior of the network, we evaluate the limiting distribution of this CTMC. The limiting or steady-state distribution Δ is defined as follows:

$$\Delta = [\Delta_1, \Delta_2, \Delta_3] \tag{7.22}$$

where

$$\Delta_j = \lim_{t\to\infty} Pr(X(t) = j) \tag{7.23}$$

For a CTMC with rate matrix $\mathbf{V} = [v_{i,j}]$, it is calculated as follows:

$$\Delta_j v_j = \sum_{i=1}^{N} \Delta_i v_{i,j} \tag{7.24}$$

and

$$\sum_{i=1}^{N} \Delta_i = 1 \tag{7.25}$$

Therefore, for our model, we determine $[\Delta_1\Delta_2\Delta_3]$ by solving the following:

$$\mathbf{A\Delta} = \mathbf{B} \tag{7.26}$$

Table 7.1: Model parameters for case studies

Parameters	Case Study I	Case Study II	Case Study III
λ_t hour^{-1}	1/8	1/3	1/8
λ_c hour^{-1}	1/80	1/30	1/80
μ_{dc} hour^{-1}	1/6	1/6	1
μ_c hour^{-1}	12	12	12
Error	10^{-5}	10^{-5}	10^{-5}

$$\text{where } \mathbf{A} = \begin{bmatrix} \lambda_t + \lambda_c & -\mu_{dc} & -\mu_c \\ \lambda_t & -(\mu_{dc} + \lambda_c) & 0 \\ \lambda_c & 0 & -\mu_c \\ 1 & 1 & 1 \end{bmatrix} \text{ and } \mathbf{B} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 1 \end{bmatrix}$$

7.2.4 Numerical Results

For the numerical results, we considered three case studies with parameter settings as listed in Table 7.1 . In case study I, trivial failures are assumed to occur with a mean value of 8 hours in relation to the traffic pattern shifts during morning, evening, and night times, which might trigger a number of SON functions. The probability of the occurrence of critical failures is assumed to be 10 times less than that of trivial failures. Cell outage detection is normally not a straight-forward task, and it may take several hours for the detection, diagnosis, and compensation of outages. Therefore, we considered μ_{dc} to be exponentially distributed with a mean value of 6 hours. In case study I, compensation is assumed to have a mean value of 5 minutes, and it also has exponential distribution. This small recovery time makes sense only when it is assumed that the SON self-healing functions involving automated diagnosis, such as those proposed in [13, 14] will be invoked for the recovery process, otherwise, a recovery time can be significantly large. In case study II, we increased the arrival rate of misconfigurations (trivial faults) from one per 8 hours to one per 3 hours. The arrival time for critical faults is assumed to be one per 30 hours. Case study II is meant to represent densely deployed

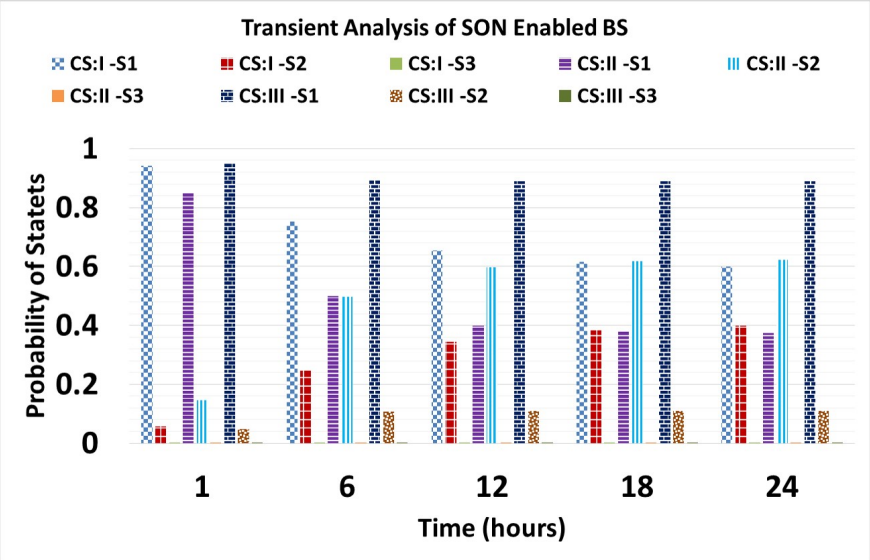


Fig. 7.6: Transient analysis of SON-enabled BS for three case studies

cells where SON functions may need to be activated and deactivated more frequently, e.g., ultra-dense mmWave-based deployment in 5G. In case study III, all parameters are assumed to be the same as those in case study I, with the exception of detection and compensation time, which is assumed to be exponentially distributed with a mean value of 1 hour. The transient analysis of the three case studies is presented in Fig. 7.6. For case study I, the probability of the BS being in an optimal healthy state is around 95% after 1 hour, and it gradually decreases to approximately 60% after a 24-hour period. There is a low probability of the BS being in an outage state, as the critical failure rate is too small in our assumed model. For case study II, the probability of the network being in a sub-optimal state is 15% after 1 hour, and it gradually increases to 62% after 24 hours, since the rate of arrival of trivial failures is high in case study II. In case study III, the probability of the BS being in an optimal state is around 88% after 24 hours. This indicates that a decreased detection and compensation time has a profound effect on network performance reliability. Therefore, the failure detection, diagnosis, and compensation time should be as small as possible to achieve maximum per-

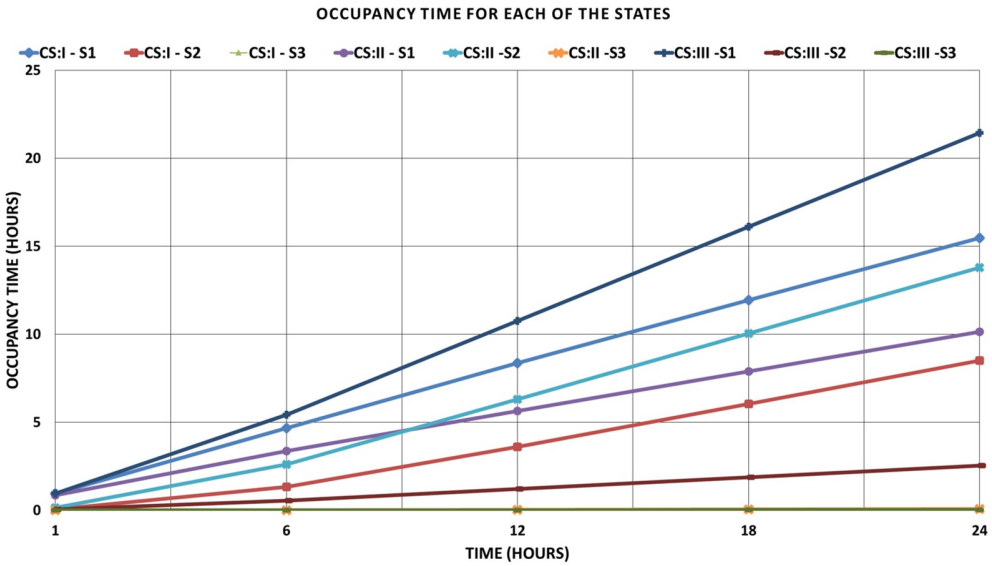


Fig. 7.7: Occupancy time of SON-enabled BS for three case studies

formance. This calls for more agile self-healing methods in emerging cellular networks where increased complexity might cause a higher fault arrival rate. The self-healing methods proposed in recent studies such as [128, 129, 133] are good candidates to overcome this problem. The occupancy time for the three case studies is illustrated in in Fig. 7.7. For case studies I and III, the network remains in an optimal state most of the time, compared to case study II in which the sub-optimal time gradually increases with the passage of time. This is a direct result of the higher rate of arrival for trivial faults in case study II. The first passage times into states 2 and 3 are portrayed in Fig. 7.8. The first passage time for the three case studies depends on the mean arrival rate of trivial as well as critical failures, so the values of λ_t and λ_c both determine when a cell first experiences degradation and complete outage. As expected, the time to the first experience of sub-optimal performance is very small, compared to complete outage. The first passage time is small in case study II, compared to the other two case studies, due to a higher rate of arrival of faults in case study II, compared to the other two case studies. The limiting or steady-state distribution is illustrated in Fig. 7.9. In the long

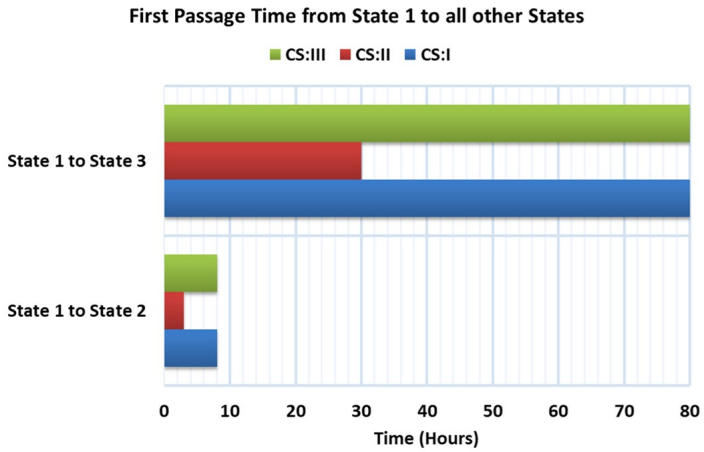


Fig. 7.8: First passage times of SON-enabled BS for three case studies

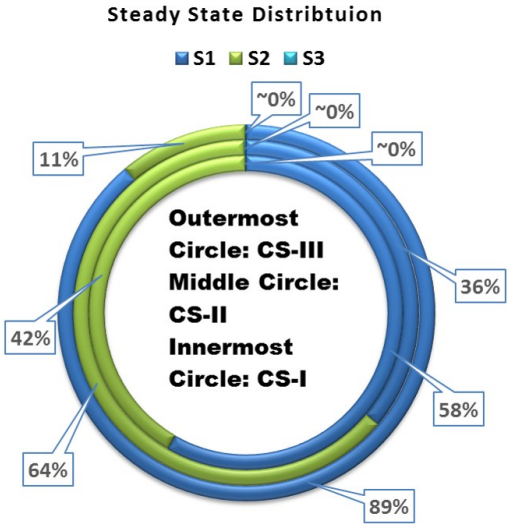


Fig. 7.9: Limiting (steady-state) distribution of SON-enabled BS for three case studies

run, a cell remains in an optimal state for 58.3% and 88.9% of the time for case studies I and III respectively. However, for case study II, it remains in that state for only 36.17% of the time (63.77% in sub-optimal state) due to a higher rate of trivial failures. The BS stays in state 3 for a negligibly small amount of time, as the critical failure rate is very small in our study.

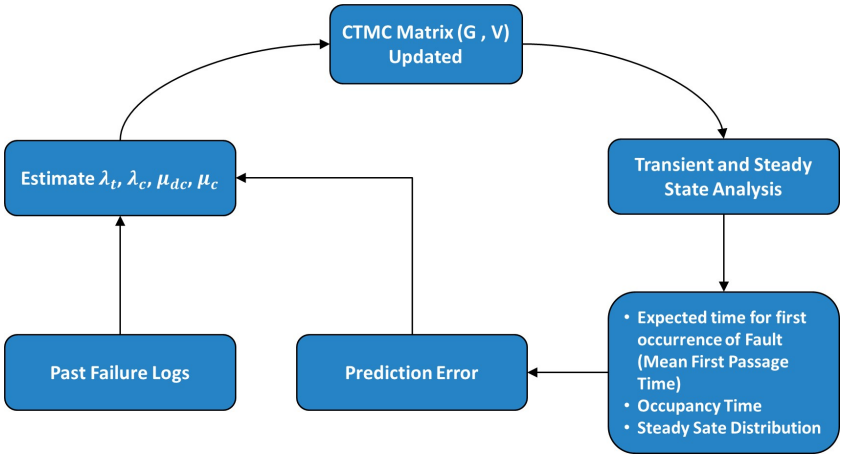


Fig. 7.10: Schematic of the proposed fault prediction framework

7.3 Fault Prediction Framework (FPF)

Utilizing the analytical model developed in section 7.2, we propose a fault predictive framework (FPF) that predicts the occurrences of faults based on a database of past failures (Fig. 7.10). The historical data related to past failures and misconfigurations of BS network parameters that occur routinely during operation of a cellular network can be utilized to estimate the λ_t, λ_c, μ_{dc}, and μ_c parameters using standard statistical tools. These estimated mean values can then be plugged into the CTMC model, and the **G** and **V** matrices can then be updated dynamically. The fitting of data to phase-type distributions has been covered in various research studies, such as in [134]. Based on updated **G** and **V** matrices, transient and steady-state analyses can then be run to compute new values for the expected time of the first occurrence of a fault, the occupancy time, and steady distribution. The difference between the predicted and actual values can be used to retrain the CTMC model parameters. In some cases, cell degradation is difficult to detect [135] e.g., in case of sleeping cells where no alarms are raised. In those cases, cell-outage/degradation detection requires expensive site visits or drive testing that may take hours or days for the sub-optimal behavior to be detected. In

the majority of cases, excessive customer complaints indicate the occurrence of faulty behavior of a cell. This results in a significant reduction in the QoS and capacity. The probability of a cell being sub-optimal at a given time period can be calculated by the proposed framework and can be exploited to minimize the degradation time. Once the predicted fault occurrence time is near, prioritizing the verification of each of the BS elements or the configuration parameters can be initiated. Similarly, the occupancy time of the BS or steady-state distribution can be used as a KPI for cell performance. If the calculated values suggest that the time period that the cell will spend in a sub-optimal or outage state is above some threshold value, then that cell can be prioritized accordingly in the optimization process. The proposed framework can also aid in the diagnosis of faults, as this is one of the most difficult tasks BS subsystem engineers face. If some record is maintained for the time interval of the occurrence of a fault and the corresponding root cause of that fault, as the expected suboptimal behavior or outage time approaches, the diagnosis should start right from the root cause already recorded in the table. This can result in a significant reduction in diagnosis time and compensation time. The CTMC model and associated FPF framework presented in this chapter can thus significantly improve the reliability of a network and provide the enhanced user experience that is expected from 5G.

7.4 Quantifying CTMC based Reliability model using Real Network Data

We also developed a CTMC-based analytical model for real historical faults data gathered from a collaborating mobile network operator. Real data can provide crucial information about the reliability of the entire network. The exploitation of a CTMC analysis is possible due to the exponential distribution

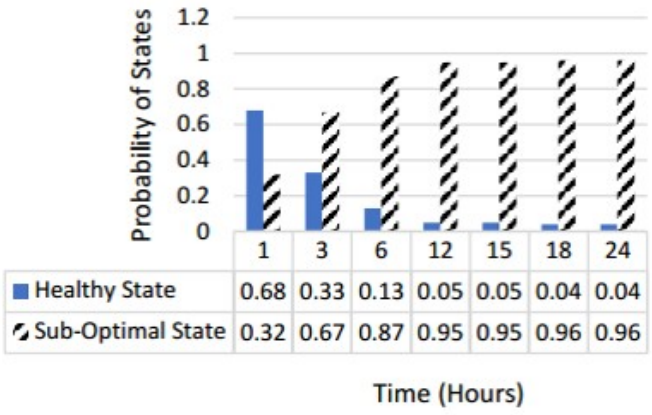


Fig. 7.11: Transient analysis for first day of network

of both the fault inter-arrival times ($1/2.589\ hour^{-1}$) and maintenance times (exponentially distributed with a mean value of 62.9 hours). For the following analysis, the network can only reside in two states: healthy and sub-optimal. Using these values, the CTMC model was developed as described in section 7.2. This analysis can be adapted each time for new values in order to compute new expected times for occupancy times, the first occurrence of a fault, and steady-state distribution.

The transient analysis using the probability matrix in Fig. 7.11 reveals that the probability of the network switching from healthy to unhealthy is very high. In fact, after 12 hours, there is a 95% chance that the system will be in an unhealthy state. After 1 day, the value for the sub-optimal state remains constant at 0.9605. However, using these probability values, a possible model can be made. For demonstration purposes, a model was developed that used a 75% threshold level to signify that a fault has occurred. The model was checked every 4 hours for a week, and the prediction accuracy was assigned a value of 1 for a correct prediction and 0 for an incorrect prediction. The model's accuracy was 27/42 or 64.29%.

The first passage time was calculated to be 2.589 hours, which is in accordance

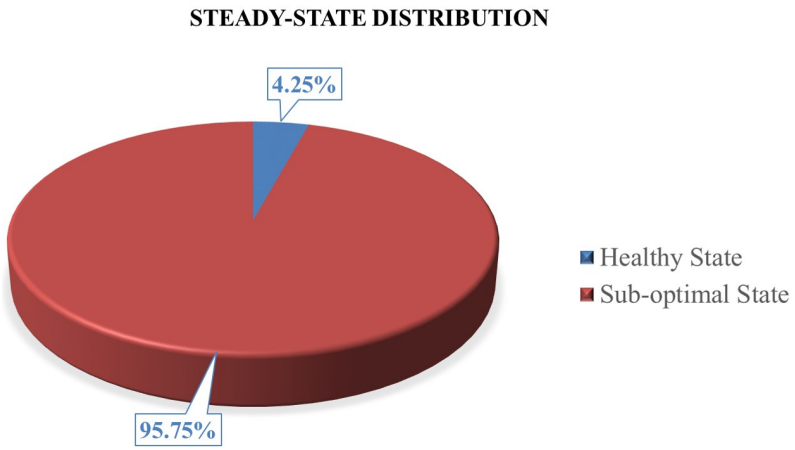


Fig. 7.12: Steady-state distribution for lifetime of network

with the mean fault inter-arrival time of 2.589 hours. This makes theoretical sense as the CMTC analysis is based on the fact that both variables follow an exponential distribution. The steady-state distribution as illustrated in Fig. 7.12, also confirms the need for a proactive approach. The distribution found that during its lifetime, the network will spend only 3.95% of its time in the healthy state, while it will spend a massive 96.05% of its time in a sub-optimal state. Such a high sojourn time in the sub-optimal state is due to the fact that we considered fault series data from multiple BSs instead of a single BS. This highlights the fact that massive densification, aimed for 5G, is consequently going to increase fault arrivals, and proactive self-healing, capable of forecasting network faults before subscribers are affected, is essential for reliable operation in 5G cellular networks.

7.5 Conclusion

In this chapter, we presented a stochastic analytical model to analyze and predict the arrival of faults on the reliability behavior of a cellular network. Assuming exponential distributions for failures and recovery, a reliability model was developed using the CTMC process. The proposed model, unlike previ-

ous studies on network reliability, is not limited to structural aspects of BSs; it takes into account diverse potential fault scenarios and is capable of predicting the expected time of the first occurrence of a fault and the long-term reliability behavior of the BS. This model can adapt itself dynamically by learning from a prior database of network failures. Three different scenarios were analyzed in terms of transient analysis, occupancy time, first passage time, and steady-state distribution. As per the numerical results, the mean arrival rate of trivial failures has a profound effect on the reliability behavior of the cellular network. Another key finding is that a substantial gain in network reliability can be achieved by reducing a BS's fault detection and recovery time, which strongly advocates the need for agile, self-healing SON functions.

CHAPTER 8

Conclusions and Future Work

> It's the possibility of having a dream come true that makes life interesting.
>
> Paulo Coelho

8.1 Conclusions

Tapping into an ultra-dense network is being widely considered as the most promising means to cope with the imminent capacity crunch. However, it is becoming clear that it is not feasible to merely rely on an increased number of cells to provide the quality of experience (QoE) expected from future cellular networks vis-à-vis 5G and beyond. The multifarious complexity and resultant resource inefficiency of such an ultra-dense, multi-tier network is another looming challenge. These above mentioned challenges call for a paradigm shift in the way cellular radio access networks architectures are designed and operated. One plausible method to address these problems is the automation of mobile cellular network (MCN) operation and optimization, dubbed self-organizing networks (SONs). While the research on SONs commenced a decade ago and is still ongoing, they may not meet the requirements of 5G in their current form, mainly for the following reasons: their reactive mode of operation, the conflict-prone design of SONs, the limited degree of freedom and lack of intelligence of the network. This dissertation addresses these limitations of state-of-the-art SONs.

To effectively tackle the spatiotemporal dynamics of network conditions, it presents a generic low-complexity framework to quantify the key facets of performance, namely, the capacity, quality of service (QoS), and power consumption of the various network topology configurations (NTCs), for enabling the SON-empowered cellular system optimization on the fly. The presented framework quantifies the multiple performance aspects of a given heterogeneous NTC through a unified set of metrics that are derived as functions of key optimization parameters, and it also presents a cross comparison of a wide range of potential NTCs. Moreover, it proposes a low-complexity heuristic approach for the holistic optimization of future heterogeneous cellular systems for joint optimality in the multiple desired performance indicators. The performance characterization framework (PCF) also provides quantitative insights into the new trade-offs involved in the optimization of emerging heterogeneous networks, and it can pave the way for much needed further research in this area.

Next, we develop and analyze a semi-Markov-based spatiotemporal mobility prediction framework for transforming a reactive SON into a proactive SON. The proposed mobility prediction model overcomes the limitation of conventional discrete-time markov chain (DTMC)-based prediction models that fail to incorporate the time dimension, i.e., "Time of next hand over (HO)." Next, we propose a novel method to map the next cell spatiotemporal HO information to the estimated future location coordinates based on the idea of Landmarks. This novel method further increases the spatial resolution of future location estimation without requiring an increase in the number of states for the semi-Markov model. The accuracy of the proposed model is quantified through experimental evaluations coupled with extensive Monte Carlo simulations.

Furthermore, it proposes a novel spatiotemporal mobility prediction-aware, proactive, sleep-mode-based energy saving (ES) optimization algorithm to solve the future 5G ultra-dense heterogeneous networks (HetNets) puzzle. The proposed AURORA framework employs an innovative concept of estimating future user locations and leveraging them to estimate future cell loads. It then devises an ES optimization problem for the estimated future network scenario. The majority of conventional reactive-style approaches are expected to dynamically solve the formulated ES problem in real time as network conditions change. However, this is close to impossible, even when substantial computing power is available. In contrast, the innovative proposed approach enables state-of-the-art heuristic techniques such as genetic algorithms (GAs) to find practically sound solutions to the formulated optimization problem predictively ahead of time. This can be an enabler for meeting ambitious 5G latency and QoS requirements. Moreover, the AURORA framework considers the interplay between the three intertwined SON functions (ES, coverage and capacity optimization [CCO], and load balancing [LB]) that exists due to the overlap between their primary optimization parameters. Therefore, it employs a co-design approach wherein the joint optimization of ON/OFF states and cell individual offset (CIO) values for small cells (SCs) does not conflict with CCO and LB objectives. Extensive simulations employing a realistic SLAW mobility model indicate that AURORA can achieve significant energy reduction gains in ultra-dense HetNets, compared to an always ON approach. A comparative performance analysis with a near-optimal performance bound indicates satisfactory robustness of the proposed AURORA framework for location estimation accuracies.

This dissertation also proposes a novel spatiotemporal mobility prediction-based, proactive LB optimization framework for HetNets by jointly optimizing

Tx power, tilts, azimuths, beam widths, and CIOs. The proposed OPERA framework solves the load-minimization optimization problem for the estimated future network scenario. The majority of conventional reactive-style approaches are expected to dynamically solve the formulated LB problem in real time as network conditions change. This is close to impossible, even when substantial computing power is available. In contrast, the innovative proposed approach—thanks to its proactiveness—allows ample time for an advanced combination of optimization heuristics namely GA and pattern search to find feasible high gain solution to the formulated optimization problem. This makes OPERA an enabler for meeting the ambitious 5G latency and QoS requirements. Moreover, the OPERA framework considers the interplay between two intertwined SON functions (LB and CCO) that exists due to the overlap between their primary optimization parameters, and it thus ensures conflict-free operation. A load-aware association strategy further bolsters the framework against location estimation accuracies. Superior performance of OPERA on several fronts compared to current schemes stems from its following features: 1) It preempts congestion instead of reacting to it; 2) it actuates more parameters than any current LB schemes thereby increasing system level capacity instead of just shifting it among cells; 3) while performing LB, OPERA simultaneously maximizes residual capacity while incorporating throughput and coverage constraints; 4) it incorporates a load aware association strategy for ensuring conflict free operation of LB and CCO SON functions.

This dissertation also presents a stochastic analytical model to analyze and predict the arrival of faults on the reliability behavior of a cellular network. Assuming exponential distributions for failures and recovery, a reliability model is developed using a continuous-time markov chain (CTMC) process.

Unlike the previous works on cellular network reliability that mostly focus on the structural aspects of cellular networks and overlook the network behavioral aspects that can cause complete or partial failures, the proposed work is more focused on developing a generic analytical model encompassing diverse fault cases such as software/hardware failures or SON conflict-attributed misconfigurations. This approach affords the flexibility to incorporate a variety of failure scenarios into the model.

On the whole, the contributions of this dissertation can make a SON more agile, intelligent, and conflict-free, and they can essentially transform it from a reactive to a proactive paradigm and hence allow it to act as a key enabler for 5G. The resource efficiency, cost saving, and service improvement achievable by the contributions of this dissertation are bound to have broad impacts on nearly every aspect of the evolving digital society that counts on cellular connectivity. Therefore, this dissertation offers key step forward towards paving the way for the commercial and technical viability of 5G and beyond.

8.2 Future Work

The PCF framework presented in chapter 3 of this dissertation requires an operator-specified policy for priority settings for various goals. This priority in itself can be devised as an optimization parameter, which can be set by a SON itself, depending on network conditions. Moreover, this priority policy can be further abstracted with a return on investment (ROI). The SON can optimize the priority values of conflicting goals to maximize the ROI for operators.

The AURORA and OPERA frameworks presented in chapter 5 and 6 respectively can be further modified with the incorporation of user-specific mobility behavior and QoS requirements of the UEs. They can be made backhaul-

aware by assigning a maximum load threshold to the cells depending on the available backhaul. In this way, core network influencing factors can be incorporated implicitly into the AURORA and OPERA frameworks. The minimum required throughputs in the problem formulations of AURORA and OPERA can be replaced with the actual spatiotemporal network usage requirements of users that can then be learned by mining call data records (CDRs) and auxiliary data sources in mobile networks; e.g., User A frequently watches cricket match. By mining CDRs and social feeds, it is possible to predict when the user A will watch an upcoming famous cricket match, and this prediction can be used for the throughput requirement of that user, instead of plain, flat, minimum required throughputs.

In addition to ES, LB, and CCO SON use cases, the presented spatiotemporal mobility prediction model in chapter 4 can also be used to transform mobility robustness optimization (MRO) SON function from reactive to proactive, since the present contribution already predicts future cells for all users (see Fig. 4.7), which in turn can be used for proactive handovers.

In the context of proactive self-healing, presented in chapter 7, the CTMC can be extended to include a non-exponential distribution for failures and recovery times, i.e., semi-Markov. Moreover, methods can be developed to efficiently estimate stochastic reliability model parameters by learning from the past failure logs collected from a real network. Also, knowledge about the future load can be exploited when the goal is to reduce the maintenance costs associated with base station (BS) power state changes. In particular, by knowing the future variations of a load, it would be easier to schedule BS power state changes in order to limit the increase in maintenance costs. Furthermore, since cell failure data are highly skewed, not enough data are available to train machine learning or statistical models for failure prediction. In this case,

synthetic cell outage data, generated through appropriate machine learning tools, can be leveraged.

Bibliography

[1] Cisco, "Cisco Visual Networking Index: Global Mobile Data Traffic Forecast, 2016-2021 - Whitepaper," Tech. Rep., 2017.

[2] I. Sobol, "Global sensitivity indices for nonlinear mathematical models and their Monte Carlo estimates," *Mathematics and Computers in Simulation*, vol. 55, no. 1, pp. 271 – 280, 2001, the Second IMACS Seminar on Monte Carlo Methods.

[3] H. Y. Lateef, A. Imran, M. A. Imran, L. Giupponi, and M. Dohler, "LTE-advanced self-organizing network conflicts and coordination algorithms," *IEEE Wireless Communications*, vol. 22, no. 3, pp. 108–117, June 2015.

[4] O. Aliu, A. Imran, M. Imran, and B. Evans, "A Survey of Self Organisation in Future Cellular Networks," *IEEE Communications Surveys Tutorials*, vol. PP, no. 99, pp. 1 –26, 2012.

[5] A. Imran and A. Zoha, "Challenges in 5G: how to empower SON with big data for enabling 5G," *Network, IEEE*, vol. 28, no. 6, pp. 27–33, Nov 2014.

[6] J. Moysen, "Self organisation for 4g/5g networks," Ph.D. dissertation, Universitat Politècnica de Catalunya (UPC), 2016.

[7] S. Haemaelaeinen, H. Sanneck, and C. Sartori, *LTE Self-Organising Networks (SON): Network Management Automation for Operational Efficiency*. Wiley,ISBN 978-1-1199-7067-5, 2012.

[8] J. Ramiro and K. Hamied, *Self-Organizing Networks (SON): Self-Planning, Self-Optimization and Self-Healing for GSM, UMTS and LTE*, 1st ed. Wiley Publishing, 2012.

[9] 3GPP, "Universal Mobile Telecommunications System (UMTS); LTE; Universal Terrestrial Radio Access (UTRA) and Evolved Universal Terrestrial Radio Access (E-UTRA); Radio measurement collection for Minimization of Drive Tests (MDT); Overall description; Stage 2," Tech. Rep., 2011-04.

[10] A. Boustani, S. Khorsandi, R. Danesfahani, and N. Mirmotahhary, "An Efficient Frequency Reuse Scheme by Cell Sectorization in OFDMA Based Wireless Networks," in *Fourth International Conference on Computer Sciences and Convergence Information Technology,. (ICCIT)*, nov. 2009, pp. 800 –805.

[11] A. Imran, E. Yaacoub, Z. Dawy, and A. Abu-Dayya, "On the capacity and spatial fairness trade-off in planning sectorization and frequency reuse," in *International Conference on Communications and Information Technology (ICCIT)*, 2013, pp. 406–410.

[12] *Physical Layer Aspects for Evolved Universal Terrestrial Radio Access (UTRA), Technical Report TR 25.814,*, 3GPP Std.

[13] T. Alsedairy, Y. Qi, A. Imran, M. A. Imran, and B. Evans, "Self organising cloud cells: a resource efficient network densification strategy," *Transactions on Emerging Telecommunications Technologies*, vol. 26, no. 8, pp. 1096–1107, 2015.

[14] A. Imran and R. Tafazolli, "Evaluation and comparison of capacities and costs of multihop cellular networks," *in 16th International Conference on Telecommunications (ICT)*, pp. 160–165, 2009.

[15] F. Velez, M. Nazir, A. Aghvami, O. Holland, and D. Robalo, "Cost/Revenue Tradeoff in the Optimization of Fixed WiMAX Deployment With Relays," *IEEE Transactions on Vehicular Technology,*, vol. 60, no. 1, pp. 298 –312, jan. 2011.

[16] G. Koutitas, A. Karousos, and L. Tassiulas, "Deployment Strategies and Energy Efficiency of Cellular Networks," *IEEE Transactions on Wireless Communications*, vol. PP, no. 99, pp. 1 –12, 2012.

[17] A. Imran, E. Yaacoub, Z. Dawy, and A. Abu-Dayya, "Planning Future Cellular Networks: A Generic Framework for Performance Quantification," in *Proceedings of the 19th European Wireless Conference (EW),*, 2013, pp. 1–7.

[18] D. Bobkov, A. Zgurskiy, O. Kozlov, M. Kolomicev, A. Sakhnevich, and M. Sklyar, "3GPP LTE access network planning," in *20th International*

Crimean Conference on Microwave and Telecommunication Technology (CriMiCo), sept. 2010, pp. 433 –434.

[19] J. Gu, Y. Ruan, X. Chen, and C. Wang, "A novel traffic capacity planning methodology for LTE radio network dimensioning," in *IET International Conference on Communication Technology and Application (ICCTA),*, oct. 2011, pp. 462 –466.

[20] J. Beyer, U. Isensee, and H. Droste, "A Measurement Based Approach to Predict the MIMO Throughput of the LTE Downlink in RF Planning Tools," in *IEEE Vehicular Technology Conference (VTC Fall),*, sept. 2011, pp. 1 –5.

[21] W. Guo and T. O'Farrell, "Relay Deployment in Cellular Networks: Planning and Optimization," *IEEE Journal on Selected Areas in Communications,*, vol. PP, no. 99, pp. 1 –10, 2012.

[22] F. Gordejuela-Sanchez and J. Zhang, "LTE Access Network Planning and Optimization: A Service-Oriented and Technology-Specific Perspective," in *IEEE Global Telecommunications Conference, (GLOBECOM)*, 30 2009-dec. 4 2009, pp. 1 –5.

[23] S. Louvros, K. Aggelis, and A. Baltagiannis, "LTE cell coverage planning algorithm optimising uplink user cell throughput," in *Proceedings of the 11th International Conference on Telecommunications (ConTEL),*, june 2011, pp. 51 –58.

[24] S. Hurley, S. Allen, D. Ryan, and R. Taplin, "Modelling and planning fixed wireless networks," *Wirel. Netw.*, vol. 16, no. 3, pp. 577–592, Apr. 2010.

[25] W. El-Beaino, A. M. El-Hajj, and Z. Dawy, "A proactive approach for LTE radio network planning with green considerations," in *19th International Conference on Telecommunications (ICT)*, april 2012, pp. 1 –5.

[26] Z. Niu, S. Zhou, Y. Hua, Q. Zhang, and D. Cao, "Energy-Aware Network Planning for Wireless Cellular System with Inter-Cell Cooperation," *IEEE Transactions on Wireless Communications,*, vol. PP, no. 99, pp. 1 –12, 2012.

[27] A. Abdel Khalek, L. Al-Kanj, Z. Dawy, and G. Turkiyyah, "Optimization Models and Algorithms for Joint Uplink/Downlink UMTS Radio Network Planning With SIR-Based Power Control," *IEEE Transactions on Vehicular Technology*, vol. 60, no. 4, pp. 1612 –1625, may 2011.

[28] Z. Li and S. Li, "LTE network planning based on game theory," in *International Conference on Computer Science and Service System (CSSS)*, june 2011, pp. 3963 –3966.

[29] V. Berrocal-Plaza, M. Vega-Rodriguez, J. Gomez-Pulido, and J. Sanchez-Perez, "Artificial Bee Colony Algorithm applied to WiMAX network planning problem," in *11th International Conference on Intelligent Systems Design and Applications (ISDA)*, nov. 2011, pp. 504 –509.

[30] S.-E. Elayoubi, O. Ben Haddada, and B. Fourestie, "Performance evaluation of frequency planning schemes in OFDMA-based networks," *IEEE Transactions on Wireless Communications*, vol. 7, no. 5, pp. 1623 –1633, may 2008.

[31] A. Imran, M. Imran, A. Abu-Dayya, and R. Tafazolli, "Self Organization of Tilts in Relay Enhanced Networks: A Distributed Solution," *Wireless Communications, IEEE Transactions on*, vol. 13, no. 2, pp. 764–779, February 2014.

[32] A. Engels, M. Reyer, X. Xu, R. Mathar, J. Zhang, and H. Zhuang, "Autonomous Self-Optimization of Coverage and Capacity in LTE Cellular Networks," *Vehicular Technology, IEEE Transactions on*, vol. 62, no. 5, pp. 1989–2004, Jun. 2013.

[33] J. Zhang, C. Sun, Y. Yi, and H. Zhuang, "A hybrid framework for capacity and coverage optimization in self-organizing LTE networks," in *Personal Indoor and Mobile Radio Communications (PIMRC), 2013 IEEE 24th International Symposium on*, Sept 2013, pp. 2919–2923.

[34] X. Zhang, Y. Zhang, R. Yu, W. Wang, and M. Guizani, "Enhancing spectral-energy efficiency for LTE-advanced heterogeneous networks: a users social pattern perspective," *Wireless Communications, IEEE*, vol. 21, no. 2, pp. 10–17, April 2014.

[35] X. Xiao, X. Tao, and J. Lu, "Energy-Efficient Resource Allocation in

LTE-Based MIMO-OFDMA Systems With User Rate Constraints," *Vehicular Technology, IEEE Transactions on*, vol. 64, no. 1, pp. 185–197, Jan 2015.

[36] K. Trichias, R. Litjens, A. Tall, Z. Altman, and P. Ramachandra, "Self-optimisation of Vertical Sectorisation in a realistic LTE network," in *Networks and Communications (EuCNC), 2015 European Conference on*, June 2015, pp. 149–153.

[37] A. Imran, M. A. Imran, and R. Tafazolli, "A novel Self Organizing framework for adaptive Frequency Reuse and Deployment in future cellular networks," in *Proc. IEEE 21st Int Personal Indoor and Mobile Radio Communications Symp,2010,(PIMRC'10)*, 2010, pp. 2354–2359.

[38] E. Yaacoub and Z. Dawy, "LTE BS Placement Optimization Using Simulated Annealing in the Presence of Femtocells," in *European Wireless 2014; 20th European Wireless Conference; Proceedings of*, May 2014, pp. 1–5.

[39] S. Hurley, "Planning effective cellular mobile radio networks," *IEEE Transactions on Vehicular Technology*, vol. 51, no. 2, pp. 243 –253, mar 2002.

[40] H. Peyvandi, A. Imran, M. A. Imran, and R. Tafazolli, "A Target-Following Regime using Similarity Measure for Coverage and Capacity Optimization in Self-Organizing Cellular Networks with Hot-Spot," in *European Wireless 2014; 20th European Wireless Conference; Proceedings of*, May 2014, pp. 1–6.

[41] D. Tsilimantos, D. Kaklamani, and G. Tsoulos, "Particle swarm optimization for UMTS WCDMA network planning," in *3rd International Symposium on Wireless Pervasive Computing,(ISWPC)*, may 2008, pp. 283 –287.

[42] H. Yang, J. Wang, X. Song, Y. Yang, and M. Wang, "Wireless base stations planning based on GIS and genetic algorithms," in *9th International Conference on Geoinformatics, 2011*, june 2011, pp. 1 –5.

[43] A. Awada, B. Wegmann, I. Viering, and A. Klein, "Optimizing the Radio Network Parameters of the Long Term Evolution System Us-

ing Taguchi's Method," *IEEE Transactions on Vehicular Technology*, vol. 60, no. 8, pp. 3825 –3839, oct. 2011.

[44] M. Galota, C. Glasser, S. Reith, and H. Vollmer, "A polynomial-time approximation scheme for base station positioning in UMTS networks," in *Proceedings of the 5th international workshop on Discrete algorithms and methods for mobile computing and communications*, ser. DIALM '01. New York, NY, USA: ACM, 2001, pp. 52–59.

[45] E. Amaldi, A. Capone, F. Malucelli, and F. Signori, "UMTS radio planning: optimizing base station configuration," in *IEEE 56th Vehicular Technology Conference, (VTC Fall. 2002)*, vol. 2, 2002, pp. 768 – 772 vol.2.

[46] N. Weicker, G. Szabo, K. Weicker, and P. Widmayer, "Evolutionary multiobjective optimization for base station transmitter placement with frequency assignment," *IEEE Transactions on Evolutionary Computation,*, vol. 7, no. 2, pp. 189 – 203, april 2003.

[47] A. Imran, M. Imran, and R. Tafazolli, "Relay Station Access Link Spectral Efficiency Optimization Through SO of Macro BS Tilts," *IEEE Communications Letters*, vol. 15, pp. 1326 – 1328, 2011.

[48] I. Viering, M. Dottling, and A. Lobinger, "A Mathematical Perspective of Self-Optimizing Wireless Networks," *IEEE International Conference on Communications, (ICC '09)*, pp. 1 –6, June 2009.

[49] R. Jain, *The Art of Computer Systems Performance Analysis: Techniques for Experimental Design,Measurement,Simulation and Modeling*. Wiley, 1991.

[50] D. Pozar, "Directivity of omnidirectional antennas," *IEEE Antennas and Propagation Magazine*, vol. 35, no. 5, pp. 50–51, 1993.

[51] K. Jacobson and W. Krzymien, "System Design and Throughput Analysis for Multihop Relaying in Cellular Systems," *IEEE Transactions on Vehicular Technology*, vol. 58, no. 8, pp. 4514 –4528, Oct. 2009.

[52] Y. Bao, H. Jiang, Y. Huang, and R. Hu, "Multi-objective Optimization

of Power Control and Resource Allocation for Cognitive Wireless Networks," in *Proc. IEEE 8th/ACIS International Conference on Computer and Information Science (ICIS'09)*, 2009, pp. 70–74.

[53] C. Song, Z. Qu, N. Blumm, and A.-L. Barabasi, "Limits of Predictability in Human Mobility," *Science*, vol. 327, no. 5968, pp. 1018–1021, 2010.

[54] S. Gambs, M.-O. Killijian, and M. N. del Prado Cortez, "Next Place Prediction Using Mobility Markov Chains," in *Proceedings of the First Workshop on Measurement, Privacy, and Mobility*, ser. MPM '12. New York, NY, USA: ACM, 2012, pp. 3:1—-3:6.

[55] D. Katsaros and Y. Manolopoulos, "Prediction in wireless networks by Markov chains," *Wireless Communications, IEEE*, vol. 16, no. 2, pp. 56–64, apr 2009.

[56] N. A. Amirrudin, S. H. S. Ariffin, N. Malik, and N. E. Ghazali, "User's mobility history-based mobility prediction in LTE femtocells network," in *RF and Microwave Conference (RFM), 2013 IEEE International*, dec 2013, pp. 105–110.

[57] X. Zhou, Z. Zhao, R. Li, Y. Zhou, J. Palicot, and H. Zhang, "Human Mobility Patterns in Cellular Networks," *IEEE Communications Letters*, vol. 17, no. 10, pp. 1877–1880, Oct 2013.

[58] J.-K. Lee and J. C. Hou, "Modeling Steady-state and Transient Behaviors of User Mobility: Formulation, Analysis, and Application," in *Proceedings of the 7th ACM International Symposium on Mobile Ad Hoc Networking and Computing*, ser. MobiHoc '06. New York, NY, USA: ACM, 2006, pp. 85–96.

[59] H. Abu-Ghazaleh and A. S. Alfa, "Application of Mobility Prediction in Wireless Networks Using Markov Renewal Theory," *Vehicular Technology, IEEE Transactions on*, vol. 59, no. 2, pp. 788–802, feb 2010.

[60] H. Farooq and A. Imran, "Spatiotemporal Mobility Prediction in Proactive Self-Organizing Cellular Networks," *IEEE Communications Letters*, vol. 21, no. 2, pp. 370–373, feb 2017.

[61] I. Schumm, "Lessons Learned from Germany's 2001-2006 Labor Market Reforms," Ph.D. dissertation, Universotu of Wurzburg, 2009.

[62] G. Corradi, J. Janssen, and R. Manca, "Numerical Treatment of Homogeneous Semi-Markov Processes in Transient Cases-a Straightforward Approach," *Methodology and Computing In Applied Probability*, vol. 6, no. 2, pp. 233–246, 2004.

[63] V. Barbu and N. Limnios, "Nonparametric Estimation for Failure Rate Functions of Discrete Time semi-Markov Processes," in *Probability, Statistics and Modelling in Public Health*, M. Nikulin, D. Commenges, and C. Huber, Eds. Springer US, 2006, pp. 53–72.

[64] J. Ghosh, S. J. Philip, and C. Qiao, "Sociological Orbit Aware Location Approximation and Routing (Solar) in DTN," State Univ. of New York at Buffalo, Tech. Rep., 2005.

[65] Q. Yuan, I. Cardei, and J. Wu, "An Efficient Prediction-Based Routing in Disruption-Tolerant Networks," *IEEE Transactions on Parallel and Distributed Systems*, vol. 23, no. 1, pp. 19–31, jan 2012.

[66] M. R. Janssen, Jacques, *Semi-Markov Risk Models for Finance, Insurance and Reliability.* Springer US, 2007.

[67] 3GPP, "3rd Generation Partnership Project; Physical layer aspects for evolved universal terrestrial radio access (e-utra)," TR 25.814 V7.1.0 Release 7, Tech. Rep., 2006.

[68] M. Gorawski and K. Grochla, *Review of Mobility Models for Performance Evaluation of Wireless Networks.* Cham: Springer International Publishing, 2014, pp. 567–577.

[69] K. Lee, S. Hong, S. J. Kim, I. Rhee, and S. Chong, "SLAW: A New Mobility Model for Human Walks," in *IEEE INFOCOM 2009 - The 28th Conference on Computer Communications.* IEEE, apr 2009, pp. 855–863.

[70] A. Mohamed, O. Onireti, S. Hoseinitabatabaei, M. Imran, A. Imran, and R. Tafazolli, "Mobility prediction for handover management in cellu-

lar networks with control/data separation," in *Communications (ICC), 2015 IEEE International Conference on*, June 2015, pp. 3939–3944.

[71] I. Ashraf, F. Boccardi, and L. Ho, "SLEEP mode techniques for small cell deployments," *IEEE Communications Magazine*, vol. 49, no. 8, pp. 72–79, aug 2011.

[72] G. Auer, V. Giannini, C. Desset, I. Godor, P. Skillermark, M. Olsson, M. Imran, D. Sabella, M. Gonzalez, O. Blume, and A. Fehske, "How much energy is needed to run a wireless network?" *IEEE Wireless Communications*, vol. 18, no. 5, pp. 40–49, oct 2011.

[73] S. Buzzi, C.-L. I, T. E. Klein, H. V. Poor, C. Yang, and A. Zappone, "A Survey of Energy-Efficient Techniques for 5G Networks and Challenges Ahead," *IEEE Journal on Selected Areas in Communications*, vol. 34, no. 4, pp. 697–709, apr 2016.

[74] M. Ajmone Marsan, L. Chiaraviglio, D. Ciullo, and M. Meo, "Optimal Energy Savings in Cellular Access Networks," in *2009 IEEE International Conference on Communications Workshops*. IEEE, jun 2009, pp. 1–5.

[75] R. Litjens and L. Jorguseski, "Potential of energy-oriented network optimisation: Switching off over-capacity in off-peak hours," in *21st Annual IEEE International Symposium on Personal, Indoor and Mobile Radio Communications*. IEEE, sep 2010, pp. 1660–1664.

[76] Z. Niu, "TANGO: traffic-aware network planning and green operation," *IEEE Wireless Communications*, vol. 18, no. 5, pp. 25–29, oct 2011.

[77] 3GPP, "LTE; Evolved Universal Terrestrial Radio Access Network (E-UTRAN); Self-configuring and self-optimizing network (SON) use cases and solutions (3GPP TR 36.902 version 9.2.0 Release 9)," Tech. Rep., 2010.

[78] F. Z. Kaddour, E. Vivier, L. Mroueh, M. Pischella, and P. Martins, "Green Opportunistic and Efficient Resource Block Allocation Algorithm for LTE Uplink Networks," pp. 4537–4550, 2015.

[79] S. Wu, Z. Zeng, and H. Xia, "Load-Aware Energy Efficiency Optimization in Dense Small Cell Networks," *IEEE Communications Letters*, vol. 21, no. 2, pp. 366–369, feb 2017.

[80] R. Tao, J. Zhang, and X. Chu, "An Energy Saving Small Cell Sleeping Mechanism with Cell Expansion in Heterogeneous Networks," in *IEEE 83rd Vehicular Technology Conference (VTC Spring)*, May 2016, pp. 1–5.

[81] Y. Qu, Y. Chang, Y. Sun, and D. Yang, "Equilibrated Activating Strategy with Small Cell for Energy Saving in Heterogeneous Network," in *2014 IEEE 80th Vehicular Technology Conference (VTC2014-Fall)*, Sep 2014, pp. 1–6.

[82] A. Ebrahim and E. Alsusa, "Interference and Resource Management Through Sleep Mode Selection in Heterogeneous Networks," *IEEE Transactions on Communications*, pp. 1–1, 2016.

[83] L.-P. Tung, L.-C. Wang, and K.-S. Chen, "An interference-aware small cell on/off mechanism in hyper dense small cell networks," in *International Conference on Computing, Networking and Communications (ICNC)*, Jan 2017, pp. 767–771.

[84] Q. Wang and J. Zheng, "A Distributed base station On/Off Control Mechanism for energy efficiency of small cell networks," in *2015 IEEE International Conference on Communications (ICC)*. IEEE, jun 2015, pp. 3317–3322.

[85] I. L. C. Araujo and A. Klautau, "Traffic-aware sleep mode algorithm for 5G networks," in *2015 International Workshop on Telecommunications (IWT)*. IEEE, jun 2015, pp. 1–5.

[86] S. Samarakoon, M. Bennis, W. Saad, and M. Latva-aho, "Opportunistic sleep mode strategies in wireless small cell networks," in *IEEE International Conference on Communications (ICC)*, June 2014, pp. 2707–2712.

[87] B. Partov, D. J. Leith, and R. Razavi, "Energy-aware configuration of small cell networks," in *IEEE 25th Annual International Symposium on Personal, Indoor, and Mobile Radio Communication (PIMRC)*, Sep 2014, pp. 1403–1408.

[88] Y. Liu, H. Tian, and G. Nie, "QoS-Aware Distributed Cell Sleep Algorithm for OFDMA Small Cell Networks," in *IEEE 82nd Vehicular Technology Conference (VTC2015-Fall)*, Sep 2015, pp. 1–5.

[89] S. Zhang, N. Zhang, S. Zhou, J. Gong, Z. Niu, and X. Shen, "Energy-Aware Traffic Offloading for Green Heterogeneous Networks," *IEEE Journal on Selected Areas in Communications*, vol. 34, no. 5, pp. 1116–1129, May 2016.

[90] Y. Sun, Y. Chang, S. Song, and D. Yang, "An energy-efficiency aware sleeping strategy for dense multi-tier HetNets," in *IEEE Globecom Workshops (GC Wkshps)*, Dec 2014, pp. 1180–1185.

[91] Z. Li, D. Grace, and P. Mitchell, "Traffic-Aware Cell Management for Green Ultradense Small-Cell Networks," *IEEE Transactions on Vehicular Technology*, vol. 66, no. 3, pp. 2600–2614, mar 2017.

[92] E. Oh, K. Son, and B. Krishnamachari, "Dynamic Base Station Switching-On/Off Strategies for Green Cellular Networks," *IEEE Transactions on Wireless Communications*, vol. 12, no. 5, pp. 2126–2136, may 2013.

[93] S. Navaratnarajah, A. Saeed, M. Dianati, and M. Imran, "Energy efficiency in heterogeneous wireless access networks," *IEEE Wireless Communications*, vol. 20, no. 5, pp. 37–43, oct 2013.

[94] K. Samdanis, P. Rost, A. Maeder, M. Meo, and C. Verikoukis, *Green Communications: Principles, Concepts and Practice*. Wiley & Sons, Ltd., 2015.

[95] H. Y. Lateef, A. Imran, and A. Abu-dayya, "A framework for classification of Self-Organising network conflicts and coordination algorithms," in *2013 IEEE 24th Annual International Symposium on Personal, Indoor, and Mobile Radio Communications (PIMRC)*. IEEE, sep 2013, pp. 2898–2903.

[96] F. Cao and Z. Fan, "The tradeoff between energy efficiency and system performance of femtocell deployment," in *7th International Symposium on Wireless Communication Systems*, Sep 2010, pp. 315–319.

[97] B. Badic, T. O'Farrrell, P. Loskot, and J. He, "Energy Efficient Radio Access Architectures for Green Radio: Large versus Small Cell Size Deployment," in *IEEE 70th Vehicular Technology Conference Fall*, Sep 2009, pp. 1–5.

[98] A. J. Fehske, I. Viering, J. Voigt, C. Sartori, S. Redana, and G. P. Fettweis, "Small-Cell Self-Organizing Wireless Networks," *Proceedings of the IEEE*, vol. 102, no. 3, pp. 334–350, mar 2014.

[99] A. J. Fehske, H. Klessig, J. Voigt, and G. P. Fettweis, "Concurrent Load-Aware Adjustment of User Association and Antenna Tilts in Self-Organizing Radio Networks," *IEEE Transactions on Vehicular Technology*, vol. 62, no. 5, pp. 1974–1988, Jun 2013.

[100] S. Luke, *Essentials of Metaheuristics (Second Edition)*. Lulu, 2013.

[101] Huawei, "WhitePaper: Five Trends to Small Cell 2020," Barcelona, Tech. Rep.

[102] J. G. Andrews, S. Singh, Q. Ye, X. Lin, and H. S. Dhillon, "An overview of load balancing in hetnets: old myths and open problems," *IEEE Wireless Communications*, vol. 21, no. 2, pp. 18–25, April 2014.

[103] F. Zhou, L. Feng, P. Yu, and W. Li, "A load balancing method in downlink LTE network based on load vector minimization," in *2015 IFIP/IEEE International Symposium on Integrated Network Management (IM)*, May 2015, pp. 525–530.

[104] P. Kreuger, O. Gornerup, D. Gillblad, T. Lundborg, D. Corcoran, and A. Ermedahl, "Autonomous Load Balancing of Heterogeneous Networks," in *2015 IEEE 81st Vehicular Technology Conference (VTC Spring)*, May 2015, pp. 1–5.

[105] P. Muñoz, R. Barco, J. M. Ruiz-Avilés, I. de la Bandera, and A. Aguilar, "Fuzzy Rule-Based Reinforcement Learning for Load Balancing Techniques in Enterprise LTE Femtocells," *IEEE Transactions on Vehicular Technology*, vol. 62, no. 5, pp. 1962–1973, Jun 2013.

[106] Z. Li, H. Wang, Z. Pan, N. Liu, and X. You, "Heterogenous QoS-

guaranteed Load Balancing in 3GPP LTE Multicell Fractional Frequency Reuse Network," *Trans. Emerg. Telecommun. Technol.*, vol. 25, no. 12, pp. 1169–1183, Dec. 2014.

[107] A. Lobinger, S. Stefanski, T. Jansen, and I. Balan, "Load Balancing in Downlink LTE Self-Optimizing Networks," in *2010 IEEE 71st Vehicular Technology Conference*, May 2010, pp. 1–5.

[108] S. Nathaniel, S. H. S. Ariffin, A. Farzamnia, and A. J. Adegboyega, "Multi-criteria load balancing decision algorithm for LTE network," in *2014 4th International Conference on Engineering Technology and Technopreneuship (ICE2T)*, Aug 2014, pp. 57–62.

[109] P. Muaoz, R. Barco, and I. de la Bandera, "Optimization of load balancing using fuzzy Q-Learning for next generation wireless networks," *Expert Systems with Applications*, vol. 40, no. 4, pp. 984 – 994, 2013.

[110] M. Sheng, C. Yang, Y. Zhang, and J. Li, "Zone-Based Load Balancing in LTE Self-Optimizing Networks: A Game-Theoretic Approach," *IEEE Transactions on Vehicular Technology*, vol. 63, no. 6, pp. 2916–2925, July 2014.

[111] Q. Ye, B. Rong, Y. Chen, M. Al-Shalash, C. Caramanis, and J. G. Andrews, "User Association for Load Balancing in Heterogeneous Cellular Networks," *IEEE Transactions on Wireless Communications*, vol. 12, no. 6, pp. 2706–2716, June 2013.

[112] S. E. Elayoubi, E. Altman, M. Haddad, and Z. Altman, "A Hybrid Decision Approach for the Association Problem in Heterogeneous Networks," in *2010 Proceedings IEEE INFOCOM*, March 2010, pp. 1–5.

[113] S. Singh, H. S. Dhillon, and J. G. Andrews, "Offloading in Heterogeneous Networks: Modeling, Analysis, and Design Insights," *IEEE Transactions on Wireless Communications*, vol. 12, no. 5, pp. 2484–2497, May 2013.

[114] J. Choi, W. Lee, Y. Kim, J. Lee, and S. Kim, "Throughput Estimation Based Distributed Base Station Selection in Heterogeneous Networks," *IEEE Transactions on Wireless Communications*, vol. 14, no. 11, pp. 6137–6149, Nov 2015.

[115] I. Viering, M. Dottling, and A. Lobinger, "A Mathematical Perspective of Self-Optimizing Wireless Networks," in *2009 IEEE International Conference on Communications*, June 2009, pp. 1–6.

[116] A. Asghar, H. Farooq, and A. Imran, "A novel load-aware cell association for simultaneous network capacity and user QoS optimization in emerging HetNets," in *2017 IEEE 28th Annual International Symposium on Personal, Indoor, and Mobile Radio Communications (PIMRC)*, Oct 2017, pp. 1–7.

[117] ——, "Concurrent Optimization of Coverage, Capacity and Load Balance in HetNets through Soft and Hard Cell Association Parameters," *IEEE Transactions on Vehicular Technology*, pp. 1–1, 2018.

[118] J. Xuan, H. Jiang, and Z. Ren, "Pseudo code of genetic algorithm and multi-start strategy based simulated annealing algorithm for large scale next release problem, technical report," 2011.

[119] A. A. Solutions. Htxcw631819x000. [Online]. Available: http://66.201.95.79/~amphenolantennas/product/htxcw631819x000/

[120] Powerwave. P45-17-xlh-rr. [Online]. Available: http://raycom-w.ru/files/import_pdf/P45-17-XLH-RR.ru.pdf

[121] S.-I. Yang, D. M. Frangopol, and L. C. Neves, "Service life prediction of structural systems using lifetime functions with emphasis on bridges," *Reliability Engineering & System Safety*, vol. 86, no. 1, pp. 39 – 51, 2004.

[122] 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, September 2008.

[123] 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)*, April 2013, pp. 4730–4735.

[124] D. Tipper, C. Charnsripinyo, and H. Shin, "Survivability analysis for mobile cellular networks," in *Comm. Networks and Distributed Systems Modeling and Simulaton Conf. (CNDS Š02)*, Jan 2002.

[125] M. Guida, M. Longo, and F. Postiglione, "Performance Evaluation of IMS-Based Core Networks in Presence of Failures," in *2010 IEEE Global Telecommunications Conference GLOBECOM 2010*, Dec 2010, pp. 1–5.

[126] H. Pham, *System Software Reliability*. Springer-Verlag London, 2006.

[127] T. Osogami and M. Harchol-Balter, "Closed form solutions for mapping general distributions to quasi-minimal PH distributions," *Performance Evaluation*, vol. 63, no. 6, pp. 524 – 552, 2006, modelling Techniques and Tools for Computer Performance Evaluation.

[128] A. Zoha, A. Saeed, A. Imran, M. A. Imran, and A. Abu-Dayya, "Data-driven analytics for automated cell outage detection in Self-Organizing Networks," in *2015 11th International Conference on the Design of Reliable Communication Networks (DRCN)*, March 2015, pp. 203–210.

[129] W. Wang, Q. Liao, and Q. Zhang, "COD: A Cooperative Cell Outage Detection Architecture for Self-Organizing Femtocell Networks," *IEEE Transactions on Wireless Communications*, vol. 13, no. 11, pp. 6007–6014, Nov 2014.

[130] K. S. Trivedi, *Probability and Statistics with Reliability, Queuing and Computer Science Applications*, 2nd ed. Chichester, UK: John Wiley and Sons Ltd., 2002.

[131] V. G. Kulkarni, *Introduction to Modeling and Analysis of Stochastic Systems*, 2nd ed. Springer-Verlag New York, 2011.

[132] A. Reibman and K. Trivedi, "Numerical transient analysis of markov models," *Computers & Operations Research*, vol. 15, no. 1, pp. 19 – 36, 1988.

[133] K. Lee, H. Lee, Y. Jang, and D. Cho, "CoBRA: Cooperative Beamforming-Based Resource Allocation for Self-Healing in SON-Based

Indoor Mobile Communication System," *IEEE Transactions on Wireless Communications*, vol. 12, no. 11, pp. 5520–5528, November 2013.

[134] A. Panchenko and A. ThÃijmmler, "Efficient phase-type fitting with aggregated traffic traces," *Performance Evaluation*, vol. 64, no. 7, pp. 629 – 645, 2007.

[135] W. Wang, J. Zhang, and Q. Zhang, "Cooperative cell outage detection in Self-Organizing femtocell networks," in *2013 Proceedings IEEE INFOCOM*, April 2013, pp. 782–790.

Appendix: Pseudocode of Algorithms

Algorithm 1 : Future Location Estimation

Input: $l_k^u, \mathbb{C}_N^u, \mathbb{T}_{HO}^u, l_{\mathbb{C}_N^u}^{LM}, SojournTime_{\max}, k, k'$
Output: $l_{k+k'}^u$

for $u \in \mathbb{U}$

If sojourn time of user "u" $\geq sojourntime_{\max}$ **OR** no training sample exist for this $\mathbb{C}_N^u$ i.e., $l_{\mathbb{C}_N^u}^{LM} = \{\}$

$$l_{k+k'}^u = l_k^u$$

Else

$$l_{k+k'}^u = l_k^u + \frac{\sqrt{(x_{\mathbb{C}_N^u}^{LM} - x_k^u)^2 + (y_{\mathbb{C}_N^u}^{LM} - y_k^u)^2}}{T_{HO}^u} * k' * \frac{[l_{\mathbb{C}_N^u}^{LM} - l_k^u]}{||(l_{\mathbb{C}_N^u}^{LM} - l_k^u)||}$$

End If

End for

Algorithm 2 Genetic Algorithm

Input:

Solution space S(AURORA: $\boldsymbol{\pi^c, P^c_{CIO}}$)(OPERA: P^c_t, θ^c_{tilt} ,ϕ^c_a ,φ^c_v,φ^c_h,P^c_{CIO}),
Objective Function Ω
Max Iterations **G**,
Solution space samples per iteration **P**,
Key samples per iteration **E**,
Mutation ratio **M**.

Output:

Solution $\boldsymbol{X} = [\boldsymbol{\pi^c, P^c_{CIO}}\ ]$ for AURORA, $[P^c_t$, θ^c_{tilt} ,ϕ^c_a ,φ^c_v,φ^c_h,$P^c_{CIO}]$ for OPERA

Generate $|\boldsymbol{P}|$ sets from **S** randomly;
Generate values of Ω for each set in **P**
Save the sets in current solution space **Pop**;
for i = 1 to G **do**
    Number of elite members in **Pop** $num_{elite} = \boldsymbol{E}$;
    select the best num_{elite} solutions in **Pop** and save them in $\boldsymbol{Pop_1}$;
    Number of crossover solutions $num_{crossover} = (|\boldsymbol{P}| * num_{elite})/2$;
    for j = 1 to $num_{crossover}$ **do**
        Randomly select 2 solutions X_A and X_B from **Pop**;
        Generate X_C and X_D by one-point crossover to X_A and X_B;
        Save X_C and X_D to Pop_2;
    end for
    for j = 1 to $num_{crossover}$ **do**
        Select a solution X_j from Pop_2;
        Mutate each element of X_j at a rate **M** and generate new solution $\acute{X}_j$;
        if $\acute{X}_j$ is non-feasible **then** $\acute{X}_j$ with a feasible solution by repairing $\acute{X}_j$;
        end if
        Update X_j with $\acute{X}_j$ in $\boldsymbol{Pop_2}$;
    end for
    Update **Pop** = $\boldsymbol{Pop_1} + \boldsymbol{Pop_2}$;
end for
Return the best solution **X** in **Pop**;

Algorithm 3 Pattern Search Algorithm

Input:

Parameter space $S(P_t^c, \theta_{tilt}^c, \phi_a^c, \varphi_v^c, \varphi_h^c, P_{CIO}^c)$

Objective Function Ω

Output: Solution $\boldsymbol{X} = [\mathrm{P}_t^c, \theta_{tilt}^c, \phi_a^c, \varphi_v^c, \varphi_h^c, P_{CIO}^c]$

$k = 0$;
while $k < iteration_{max}$ **do**
    Determine a step size s_k using exploratory search algorithm;
    Test Ω at x_0 and two more points x_1 and x_2 in a triangle;
    Label best, good and worst points as x_B, x_G and x_W;
    Reflect x_W on the plane as x_R;
    if $\Omega(x_R) > \Omega(x_G)$ **then**
        if $\Omega(x_R) > \Omega(x_B)$ **then** replace x_W with x_R;
        else Find $x_E | 2x_R - (x_B + x_G)/2$, find $\Omega(x_E)$
            if $\Omega(x_E) > \Omega(x_B)$ **then** replace x_W
            end if
        end if
    else
        if $\Omega(x_R) < \Omega(x_W)$ **then** replace x_W with x_R;
            Compute $x_C = ((x_B + x_G)/2) + x_R)/2$, find $\Omega(x_C)$
        else Compute $x_C = ((x_B + x_G)/2) + x_W)/2$, find $\Omega(x_C)$
        end if
        if $\Omega(x_C) < \Omega(x_W)$ **then** replace x_W with x_C;
        else Compute $x_S = (x_B + x_W)/2$ and replace x_W with x_S and $x_G = (x_B + x_G)/2$
        end if
    end if
    Compute $p_k = \Omega(x_k) - \Omega(x_k + s_k)$
    if $p_k > 0$ **then** $x_{k+1} = x_k + s_k$
    else$x_{k+1} = x_k$
    end if
    Update Pattern Vectors and step size $k = k + 1$
end while
Return $\mathbf{X} = [P_t^c, \theta_{tilt}^c, \phi_a^c, \varphi_v^c, \varphi_h^c, P_{CIO}^c]$