CCO - LB Summary

hakn2@viettel.com.vn

May 2025

Summary of Contributions

- **Joint Optimization Framework:** A unified, multi-objective optimization framework is proposed to simultaneously improve coverage, capacity, and load balance. It models the dependencies between user throughput, signal quality, and resource availability using parameter controls like transmit power, CIO, and antenna tilt.

- **Algorithm Design:** Explores both model-free methods (e.g., genetic algorithms, reinforcement learning) and gradient-based approaches, demonstrating how each can be leveraged to optimize cell configurations under realistic constraints. The gradient-based method is implemented using a differentiable pipeline that enables efficient tuning of transmit power based on a learnable throughput model.

- **Simulation and Dataset Design:** For generalizability, synthetic user data and randomized network snapshots are used to simulate high-load conditions and assess optimization strategies.

  Experimental Validation: The proposed methods are evaluated across key metrics such as throughput distribution, outage ratio, and user-level QoE. Results demonstrate clear performance improvements, particularly in reducing outage and improving fairness across cells.

- **Foundations for Time-Aware SON Design:** We outlines a path forward toward proactive SON strategies by integrating temporal patterns into network control models. Future work is proposed to incorporate time-series forecasting, allowing the system to predict and adapt to daily traffic dynamics.

Chapter 1

Self-Organizing Networks (SON) Optimization Objectives

Our goal is to solve Coverage, Capacity, and Load Balancing optimisation challange (CCO-LB) for 4G-LTE networks, through adjustment of cell parameters: Antenna Tilt, Power, and Cell Individual Offset (CIO).

1.1 Coverage Optimization

Goal: Ensure that users experience adequate signal strength throughout the network to reduce coverage holes and dead zones.

Importance: Poor coverage results in an inability to connect to the network or frequent call drops.

Quantitative Metrics:

- **RSRP (Reference Signal Received Power):** Measures the signal power received from a cell. Good coverage typically requires RSRP above a threshold, such as -100 dBm.
- **Coverage Probability:**

$$\text{Coverage Probability} = \frac{\text{\# of users with RSRP above threshold}}{\text{Total \# of users}}$$

- **Percentage of Area or Users Above Threshold:** Indicates the fraction of the geographical area or users with signal above a defined threshold.

1.2 Capacity (Throughput) Optimization

Goal: Maximize the data-carrying capability of the network to ensure high user and system throughput.

Importance: High capacity is essential for services requiring fast data transfer such as video streaming, downloads, and online gaming.

Quantitative Metrics:

- **Average User Throughput:**

$$\text{Average Throughput} = \frac{\text{Sum of user throughputs}}{\text{Number of users}}$$

- **Cell Throughput:** Total throughput for each cell, calculated as the sum of all user throughputs within that cell.
- **5th/50th/95th Percentile Throughput:** Describes worst-case, median, and near-best-case user experiences.
- **Spectral Efficiency:**

$$\text{Spectral Efficiency} = \frac{\text{Throughput (bps)}}{\text{Bandwidth (Hz)}}$$

1.3 Load Balancing Optimization

Goal: Distribute traffic evenly across network cells to prevent overloading certain cells while others remain underutilized.

Importance: Balanced load ensures better user experience and efficient resource utilization across the network.

Quantitative Metrics:

- **Cell Load:**

$$\text{Load} = \frac{\text{Resources used}}{\text{Total available resources}}$$

this is precisely the way we are calculating cell load in our system of: Ericsson 4G LTE network:

$$\text{Cell Load (TU PRB)} = \frac{100 \times (\sum \text{pmPrbUsedDlSum} \,/\, \sum \text{pmPrbUsedDlSamp})}{(\sum \text{pmPrbAvailDl} \,/\, 60 \,/\, 60000)}$$

Explanation: This equation estimates the percentage of PRB usage over time in the downlink direction. The numerator represents the average PRB utilization per sample, and the denominator normalizes the total available PRBs over time.

- **Load Variance Across Cells:**

$$\text{Load Variance} = \text{Var}(\text{Load across all cells})$$

- **Gini Coefficient for Load:** Measures inequality of load distribution across cells (0 = perfect equality, 1 = complete inequality).

Objective	Metric Examples	Interpretation
Coverage	% users with RSRP > threshold	Higher is better
Capacity	Avg throughput, 5th percentile throughput	Higher is better
Load Balancing	Load variance, Gini coefficient	Lower is better

Table 1.1: Summary of SON Optimization Objectives and Their Metrics

1.4 Coverage, Capacity, and Load Balancing Interdependence

Coverage ensures that users are within the radio footprint of a base station, but strong signal strength alone does not guarantee high data rates. **Capacity** defines the data-carrying capability of the network and depends heavily on how **radio resources are allocated** among users. If a cell is **overloaded**, even users with high SINR (Signal-to-Interference-plus-Noise Ratio) will receive a **smaller share of available resources**, which reduces their effective **throughput**.

Load balancing helps mitigate this by steering users away from heavily loaded cells and into under-loaded neighboring cells. This increases the **available resource share per user** and leads to improvements in both average and worst-case throughput.

In systems like LTE and 5G, downlink data is transmitted in units called **resource blocks** (RBs). Assuming that each cell j has a total of N_{RB} resource blocks available per Transmission Time Interval (TTI), and that these are evenly shared among α_j active users, the throughput of user i associated with cell j can be approximated as [1]:

$$\text{Throughput}_i = \frac{N_{\text{RB}} \cdot W_{\text{RB}}}{\alpha_j} \cdot \eta_i \tag{1.1}$$

where:

- N_{RB} is the total number of RBs in cell j,
- W_{RB} is the bandwidth per RB (e.g., 180 kHz in LTE),
- α_j is the number of active users in the cell,
- η_i is the spectral efficiency of user i (in bits/s/Hz), which is a function of SINR.

This equation shows that:

- Throughput decreases as load α_j increases, due to reduced resource allocation per user.
- Throughput increases with spectral efficiency η_i, which depends on channel quality.
- Throughput scales linearly with the number and bandwidth of resource blocks.

[1] 3rd Generation Partnership Project, Self-configuring and self-optimizing network (SON) use cases and solutions, 3GPP TR 36.902, Rev. V9.2.0 Release 9, 2010.

This relationship justifies the need for *joint optimization*: balancing load across cells ensures that more users receive an adequate share of radio resources, leading to fairer and more efficient capacity utilization across the network.

1.5 Challenges in Joint Parameter Optimization for SON

While optimizing coverage, capacity, and load balancing independently provides some performance improvements, their strong interdependence makes joint optimization a far more effective—yet significantly more complex—approach. In practice, these objectives are coupled through shared radio resources, interference behavior, and user mobility patterns. The task becomes more challenging when adjusting multiple interdependent parameters, particularly **Transmit Power**, **Antenna Tilt**, and **Cell Individual Offset (CIO)**.

These parameters are critical levers of network behavior:

- **Transmit Power** (P_j) directly impacts the signal strength received by users, affects cell range, and modifies the interference perceived by neighboring cells.
- **Antenna Tilt** (θ_j)—especially in the vertical domain—can focus or widen the cell coverage area, thereby influencing both the signal footprint and potential overlap with other cells.
- **Cell Individual Offset (CIO)** modifies user handover behavior by virtually biasing users toward or away from certain cells, helping with load balancing.

However, tuning these parameters in isolation can lead to suboptimal or even degraded network performance due to their intrinsic interactions. For instance:

- Increasing the transmit power of one cell improves its coverage but may increase interference for neighboring cells.
- Adjusting antenna tilt can shift traffic across coverage zones, potentially overloading certain cells while leaving others underutilized.
- CIO adjustments affect user-cell association patterns and, indirectly, the experienced interference and traffic distribution.

Optimization Complexity

The joint optimization problem is inherently non-convex and high-dimensional, especially in dense heterogeneous networks with many tunable parameters across a large number of cells. Furthermore, the objective function must reflect end-user experience and consider:

- SINR-dependent throughput models,
- Time-varying user distribution and mobility,

- Load-aware cell association dynamics.

In the next chapter we survey main approaches to solve this challenge, and explain in detail the methods we are working on.

Chapter 2

Comparison of Model-Free and GD-Based Approaches

2.1 Introduction

Given a network state with users, cells, and signal measurements X, we aim to learn a function f that is able to find the optimal set of parameters y that maximise/minimise an objective:

$$f_\theta \colon \mathcal{X} \to R^3, \quad \mathbf{y}_j = f_\theta(\mathcal{X}) = [P_j, \mathrm{CIO}_j, T_j], \quad \forall j \in \mathcal{C}$$

where P_j, CIO_j, and T_j are transmit power, cell individual offset, and tilt for cell j, respectively.

In optimizing complex systems like cellular networks, two main methods are used: model-free and gradient descent (GD)-based approaches. Each has unique strengths and weaknesses.

2.2 Model-Free Approaches

Model-free approaches optimize without explicit system models, relying on empirical data. They learn f through iterative feedback from the environment. The structure of f is determined during training, allow greater flexibility. Common model-free techniques include:

- **Reinforcement Learning (RL):** Learns optimal actions based on rewards from the environment.
- **Bayesian Optimization (BO):** Uses a probabilistic model to efficiently explore configurations.
- **Meta-Heuristic Methods:** Techniques like genetic algorithms search for optimal solutions without a model.

Pros and Cons

Pros	Cons
Flexible and adaptable	Scalability issues in high dimensions
Simple to implement	Sample inefficiency due to extensive probing

Table 2.1: Pros and Cons of Model-Free Approaches

2.3 GD-Based Approaches

In a model-based approach, you start by defining a specific functional form for f, often based on prior knowledge or assumptions about the relationship between parameters y and environmental conditions X; then explicitly choose a loss function (can be least squares, maximum likelihood estimation, etc.) that reflects the objective (maximize/minimize). GD-based approaches rely on a differentiable model to optimize parameters y using gradient calculations. They provide efficient updates but require accurate models.

Common techniques include:

- **Gradient Descent:** Updates parameters based on their impact on the objective function.

Pros and Cons

Pros	Cons
Fast convergence in complex landscapes	Needs an accurate model
Immediate feedback from gradients	Risk of local minima

Table 2.2: Pros and Cons of GD-Based Approaches

2.4 Conclusion

Both approaches mandates a suitable objective function, either in the form of a reward or loss function, representing the end-user performance, while taking cell load and network parameter dependencies into account.

- Choose **Model-Based** or **GD-based** when you have strong domain knowledge and need interpretability.
- Opt for **Model-Free** when dealing with complex, dynamic environments where adaptability is crucial.

2.5 Current Approaches

In this section, we present distinct techniques that we have emphasized for addressing the functions of CCO and LB within Self-Organizing Networks (SON). Each of these methods exemplifies the primary techniques discussed previously. The three approaches we have explored are:

- **Genetic Algorithm** (Model-free): This method employs principles of natural selection to iteratively improve solutions.
- **Reinforcement Learning** (Model-free): This approach utilizes feedback from the environment to optimize decision-making processes over time.
- **Gradient Descent** (GD-based): This technique systematically minimizes a cost function by iteratively adjusting parameters based on the gradient. We are currently utilise this method in our system.

2.5.1 Genetic Algorithm (Model-free)

As mentioned in the previous sections, we always mandate for a suitable objective function. The authors in [1] defined the objective function as the geometric mean of user throughputs across users and across cells:

$$\max_{P_t^c, \psi_{\text{tilt}}^c, P_{CIO}^c} \left(\prod_{c \in C} \left(\prod_{u \in U_c} \omega_u^c \log_2(1 + \gamma_c^u) \right)^{1/|U_c|} \right)^{1/|C|} \tag{2.1}$$

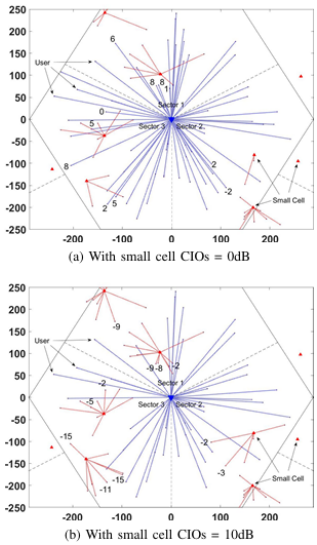
Subject to:

$$P_{t,\min} \leq P_t^c \leq P_{t,\max} \, \forall c \tag{2.2}$$
$$\psi_{\min} \leq \psi_{\text{tilt}}^c \leq \psi_{\max} \, \forall c \tag{2.3}$$
$$P_{CIO,\min} \leq P_{CIO}^c \leq P_{CIO,\max} \, \forall c \tag{2.4}$$
$$\frac{1}{|C|} \sum_{c \in C} \frac{1}{|U_c|} \sum_{u \in U_c} 1(P_{r,u}^c \geq P_{th}^c) \geq \bar{\omega} \tag{2.5}$$
$$\tau_u \geq \hat{\tau}_u, \, \forall u \tag{2.6}$$
$$\eta_c < 1, \, \forall c \in C \tag{2.7}$$

This objective function naturally embed coverage, capacity, and load balancing:

- The inner geometric mean across users U_c inside a cell ensures fairness: A cell's performance is bad if even one user has very low throughput (coverage + capacity at user level).
- The outer geometric mean across cells C encourages balanced performance across cells (load balancing).
- Naturally include capacity by maximizing overall throughput.
- Naturally include coverage by ensuring minimum RSRP thresholds.

[1] Asghar, Ahmad, Hasan Farooq, and Ali Imran. "On concurrent optimization of coverage, capacity and load balance in HetNets through joint self-organization of soft and hard cell association parameters." IEEE Transactions on Vehicular Technology 67.9 (2018): 8781-8795.



(a) With small cell CIOs = 0dB

(b) With small cell CIOs = 10dB

Figure 2.1: Empirical evidence show that during optimisation , an user-association method that does not considers user traffic demands and current cell loads can cause significant drop in signal quality (SINR). The negative influence of degraded SINR on user throughput can be partially offset if new cell can allocate enough surplus PRBs compared to old cell and thus satisfy required QoE.

- Include load balancing because the geometric mean punishes overloaded cells (where users' throughputs are low).

The cell load η_c is defined as the ratio of PRBs occupied in cell during a Transmission Time Interval (TTI) and total PRBs available in the cell:

$$\eta_c = \frac{1}{N_b^c} \sum_{u \in \mathcal{U}_c} \frac{\hat{\tau}_u}{\omega_B \cdot \log_2(1 + \gamma_u^c)} \tag{2.8}$$

Furthermore, to prevent the negative impact of sudden drop in SINR on throughput during each time we re-assign users to cell in each GA optimisation iteration (see Fig.2.1), a Load-aware User Association method is also implemented.

To solve the complex and non-convex problem like (2.1), the Genetic Algorithm is known to be one of the most suitable heuristic algorithms. It evolves a population of candidate solutions using *crossover* and *mutation*, making it well suited for complex, multivariable problems by promoting global exploration and avoiding local optima. The operation of GA is depicted in the Fig.2.2

We can apply directly the proposed methods in our pipeline, as showed in the Fig.2.3

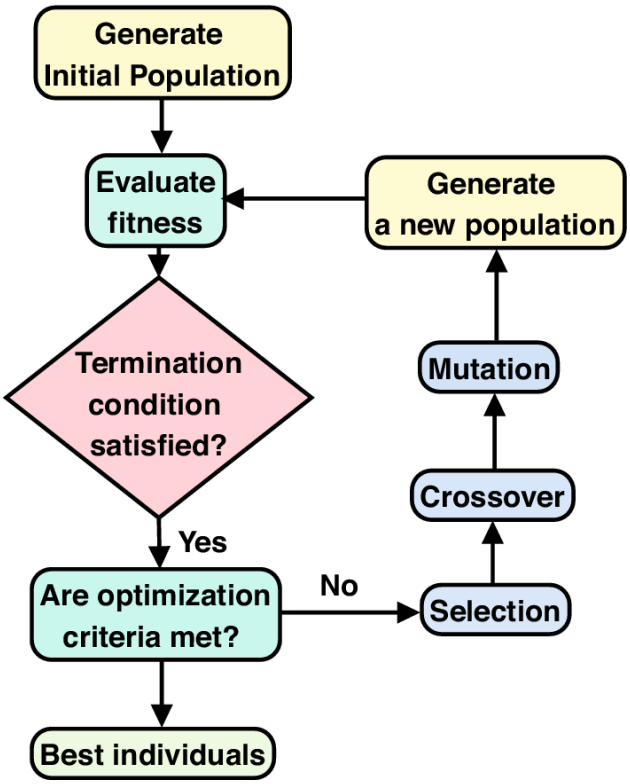


Figure 2.2: Genetic Algorithm Operation

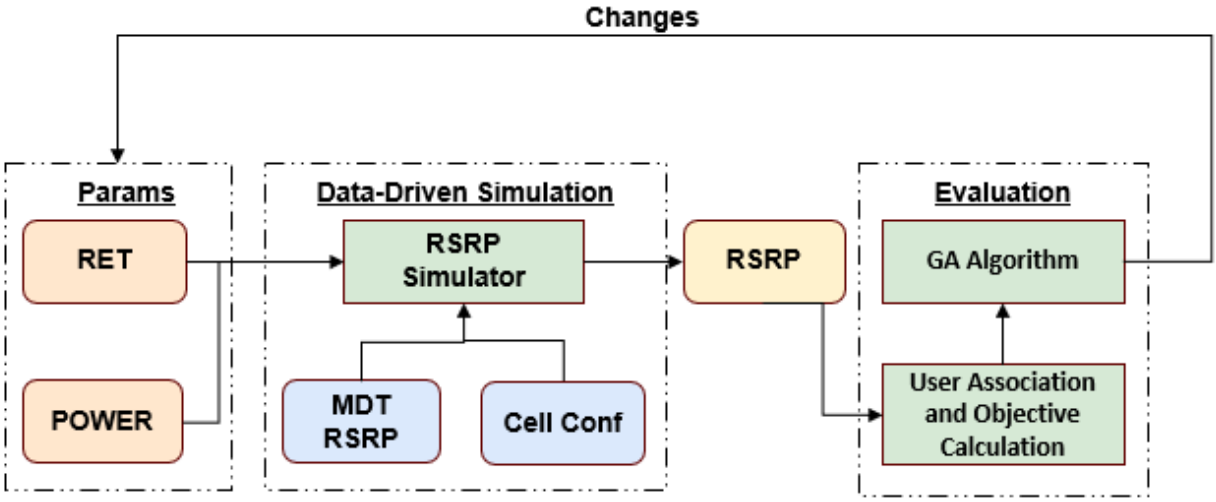


Figure 2.3: Framework for Tilt and Power Adjustment Utilise Genetic Algorithm

2.6 Gradient-Based Optimization

Author in [2] have a novelty approach to the optimisation problem, by modeling the entire network pipeline as a differentiable function. This enables the use of gradient-based optimization (e.g., Adam optimizer) to directly adjust transmit powers in a way that improves network performance. Their approach is also a definition of a throughput function, that is embedded with cell load element

$$t_i = B_{RB} \cdot N_{RB} \cdot \sum_{j=1}^{C} \frac{a_{i,j} \cdot \eta(s_{i,j})}{1 + \alpha(u_j)} \tag{2.9}$$

, and a loss function that can be optimised through gradient calculation

$$\text{Loss}_{\text{outage}} = \frac{1}{U} \sum_{i=1}^{U} \sigma(t^{\text{thresh}} - t_i) \tag{2.10}$$

The pipeline to appoximate and maximise the throughput objective function is depicted in Fig.2.4.

The optimization pipeline implemented follows these steps:

- **Soft User Association:** Instead of assigning each user to a single best cell, a softmax function is used to compute association probabilities. This makes the assignment differentiable.
- **Load Estimation:** The expected load on each cell is estimated from the user association probabilities. A non-linear, differentiable function (learned from data) maps the number of users to a cell load value.
- **SINR Calculation:** SINR for each user is computed based on the power received from the serving cell, the interference from all other cells, and thermal noise. The interference is weighted by the estimated load on each cell.
- **Spectral Efficiency Mapping:** The SINR is converted into a spectral efficiency value using a differentiable approximation of the 3GPP CQI-to-efficiency lookup table.
- **Throughput Computation:** Each user's throughput is calculated based on the shared bandwidth in its associated cell and its spectral efficiency.
- **Objective Function:** The objective is to minimize the outage loss, defined as the fraction of users whose throughput falls below a given threshold. This is modeled using a smooth approximation (sigmoid function), enabling gradient-based optimization.
- **Gradient-Based Optimization:** The transmit power for each cell is treated as a learnable parameter. Gradients of the loss with respect to power values are computed via automatic differentiation, and parameters are updated using the Adam optimizer.

[2] Eller, Lukas, Philipp Svoboda, and Markus Rupp. "A differentiable throughput model for load-aware cellular network optimization through gradient descent." IEEE Access 12 (2024): 14547-14562.

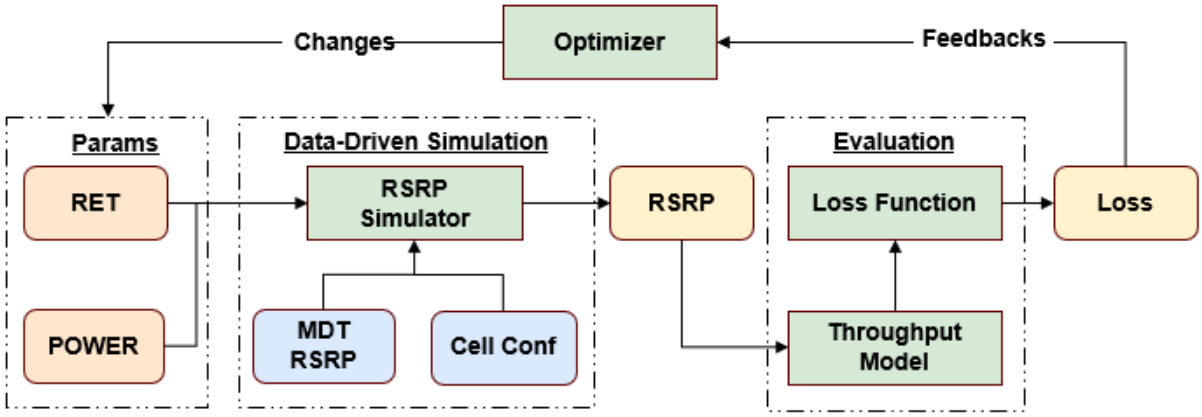


Figure 2.4: Differentiable Framework for Tilt and Power Adjustment to Optimise CCO-LB

Assuming predicted parameters shift RSRP values additively:

$$r_{ij}^{\text{new}} = r_{ij}^{\text{base}} + \Delta P_j + \Delta T_j + \Delta \text{CIO}_j$$

Chapter 3

Experiments

This section presents the experimental setup and methodology used to evaluate the 2 proposed methods, genetic algorithm and differentiable framework for cellular network optimization.

3.1 Dataset

We used a real-world dataset from Hai Hau, Nam Dinh province, to evaluate our proposed algorithms. This data set includes information on cell connections and user activity in different time periods from March 25, 2025, to April 1, 2025. We focus on peak hours for evaluation, as these times may best represent the network conditions where cell load optimization is critical. However, our analysis revealed that even during peak usage recorded, there were 2,254 users in 858 cells. This situation does not represent a scenario in which optimizing cell load is needed. Consequently, for simplification, we first opted to generate user data at random.

3.1.1 Simulation Setup

We simulate a single-tier cellular network deployed over a square area of 1000 $\times$ 1000 meters. A total of 25 base stations (cells) and 1500 user devices are randomly distributed within this area (See an example of this setup in Fig.). Each base station operates over a 20 MHz bandwidth, subdivided into 100 Physical Resource Blocks (PRBs) of 180 kHz each. The reference transmit power is set to 43 dBm, and each base station is allowed to adjust its transmit power within $\pm$15 dB of this reference level. RSRP is given by the difference between the transmit power of the base station and the path loss.

Path loss between users and cells is modeled using a log-distance path loss model with a path loss exponent of 3.5 (urban setting). Noise power is computed using the standard formula with a thermal noise floor of 174 dBm/Hz, a 10 MHz bandwidth, and a noise figure of 5 dB: $-174 + 10 * log10(bandwidth) + noise\_figure$.

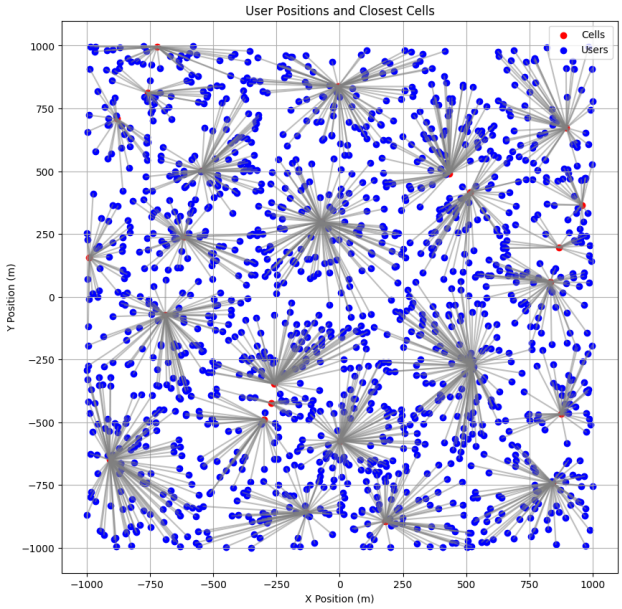


Figure 3.1: An example of random network snapshot.

3.2 Performance Evaluation and Result

3.2.1 Differentiable Framework

We run the framework through the dataset we had generated above. The optimization is run over multiple simulated network snapshots. In each case, transmit power values are initialized and updated over 100 optimization steps. Performance is tracked in terms of average user throughput, UE shared throughput distribution, and outage percentage. Results showed in Fig.3.2, Fig.3.3, and Fig.3.4 show better network performance after running the pipeline.

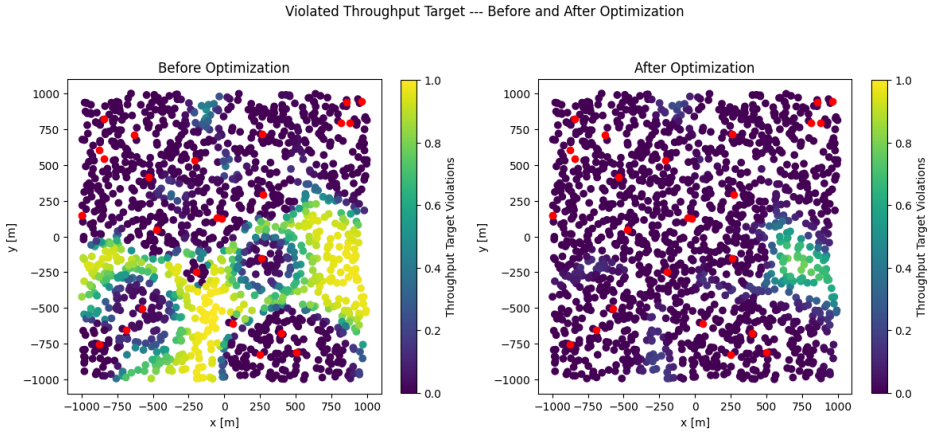


Figure 3.2: Outage ratio before and after optimisation. After optimisation, more UE can achieve expected rate.

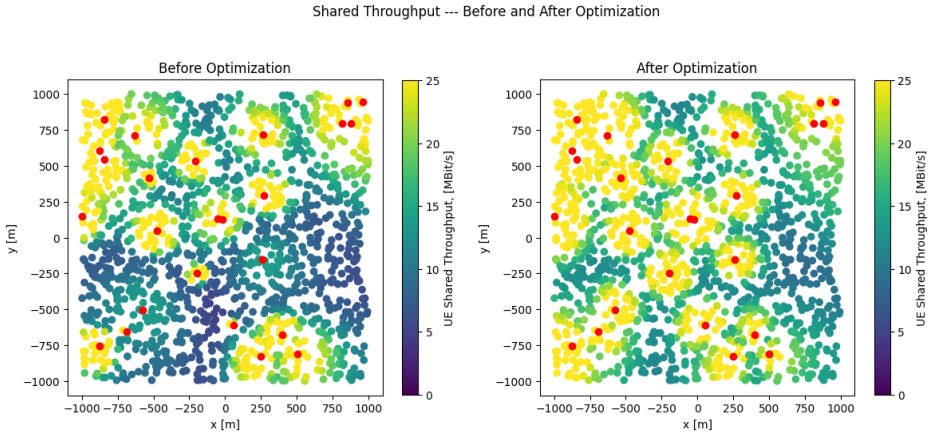


Figure 3.3: UE shared throughput before and after optimisation. After optimisation, overall throughput of users increase since users in congested cells are assigned to their neighbors.

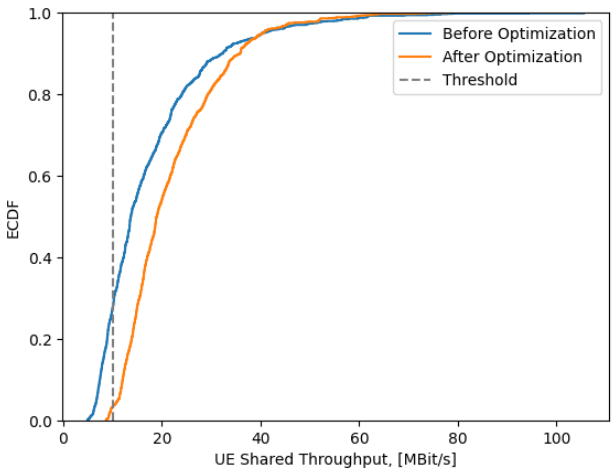


Figure 3.4: Throughput CDF before and after optimisation. Since more users are received higher rate, the distribution will shift to the right.

Chapter 4

Future Works

The current classical SON, such as MLB, is reactive when a problem occurs. It waits to observe and spot the problem (for example, most of the LB solutions are reactive and are designed to perform LB dynamically in real-time after observing the congestion). To meet 5G real time requirements, the SON mode has to be transformed to a proactive one (see Fig.4.1) so as to observe the situation and predict the problem before it occurs[1].

In dynamic cellular networks, user distribution, interference, and traffic demand vary over time. The analysis on the Hai Hau-Nam Dinh dataset, in spite of its small size, still give us a sense that network behaviour can be profiling and is able to be predicted. To enable time-aware decision making, we can incorporate temporal information into the model using embedded timestamp features. In the future, we will investigate time-series forecasting in the network traffic to supports time-aware decision making to predict network behaviors during the day.

[1]Fourati, Hasna, et al. "Comprehensive survey on self-organizing cellular network approaches applied to 5G networks." Computer Networks 199 (2021): 108435.

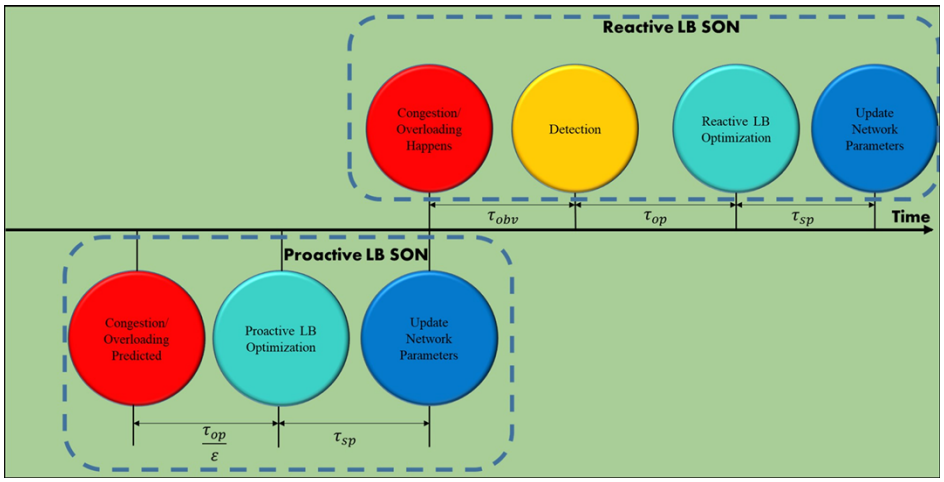


Figure 4.1: Time line diagram for Proactive and Reactive LB SON functions

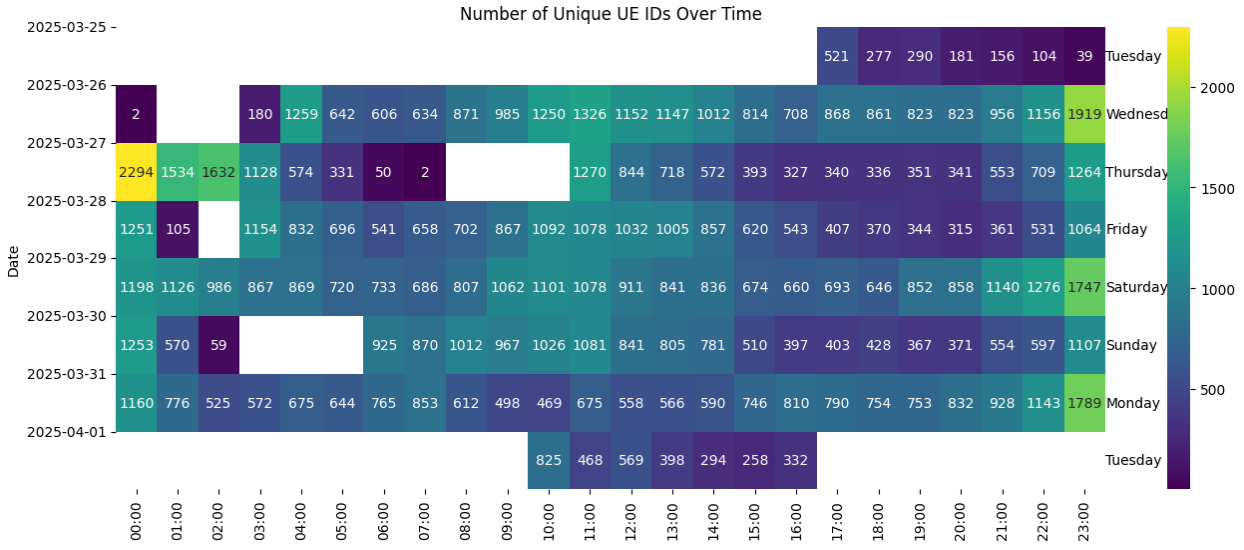


Figure 4.2: An example of traffic pattern from our real-world small dataset. The network seems busiest from 23:00 to 01:00.