MicroOpt: Model-driven Slice Resource Optimization in 5G and Beyond Networks

Muhammad Sulaiman, Mahdieh Ahmadi, Bo Sun, Mohammad A. Salahuddin, Raouf Boutaba, Aladdin Saleh

ROGERS

UNIVERSITY OF WATERLOO
DAVID R. CHERITON SCHOOL OF COMPUTER SCIENCE

RESEARCH PROBLEM

BACKGROUND

- **5G Network:** Comprises Virtual Network Functions (VNFs) within the Radio Access Network (RAN), Transport, and Core
- **Network Slicing:** Involves creating isolated, virtual networks tailored to specific use-cases (e.g., eMBB slices tailored of VR/XR, URLLC slice for remote driving, telesurgery and mMTC for IOT applications)
- **Slice QoS & SLA:** Service Level Agreements (SLAs) define the minimum Quality of Service (QoS) 5G slices must receive
- **Resource Allocation vs. QoS:** The amount of resources allocated to VNFs directly impacts the slice's QoS

MOTIVATION

- **Dynamic Traffic Patterns:** Slices experience time-varying traffic
- **Resource Efficiency:** Resource allocation for peak traffic leads to resource over-provisioning

PROBLEM STATEMENT & CHALLENGES

- **Resource Optimization:** Minimize resource allocation while meeting QoS requirements
  - **Network Complexity:** Accurately modeling complex, multi-VNF slices
  - **Real-Time Adaptation:** Fast, efficient resource adjustment to varying slice traffic

SOLUTION

- **ML-Based VNF & Slice Modeling:** Leveraging machine learning for modeling VNFs that can be composed to create end-to-end 5G slice models
- **Resource Allocation Algorithm:** Using Lagraingian primal-dual algorithm coupled with gradient descent for fast, near-optimal resource scaling under QoS constraints

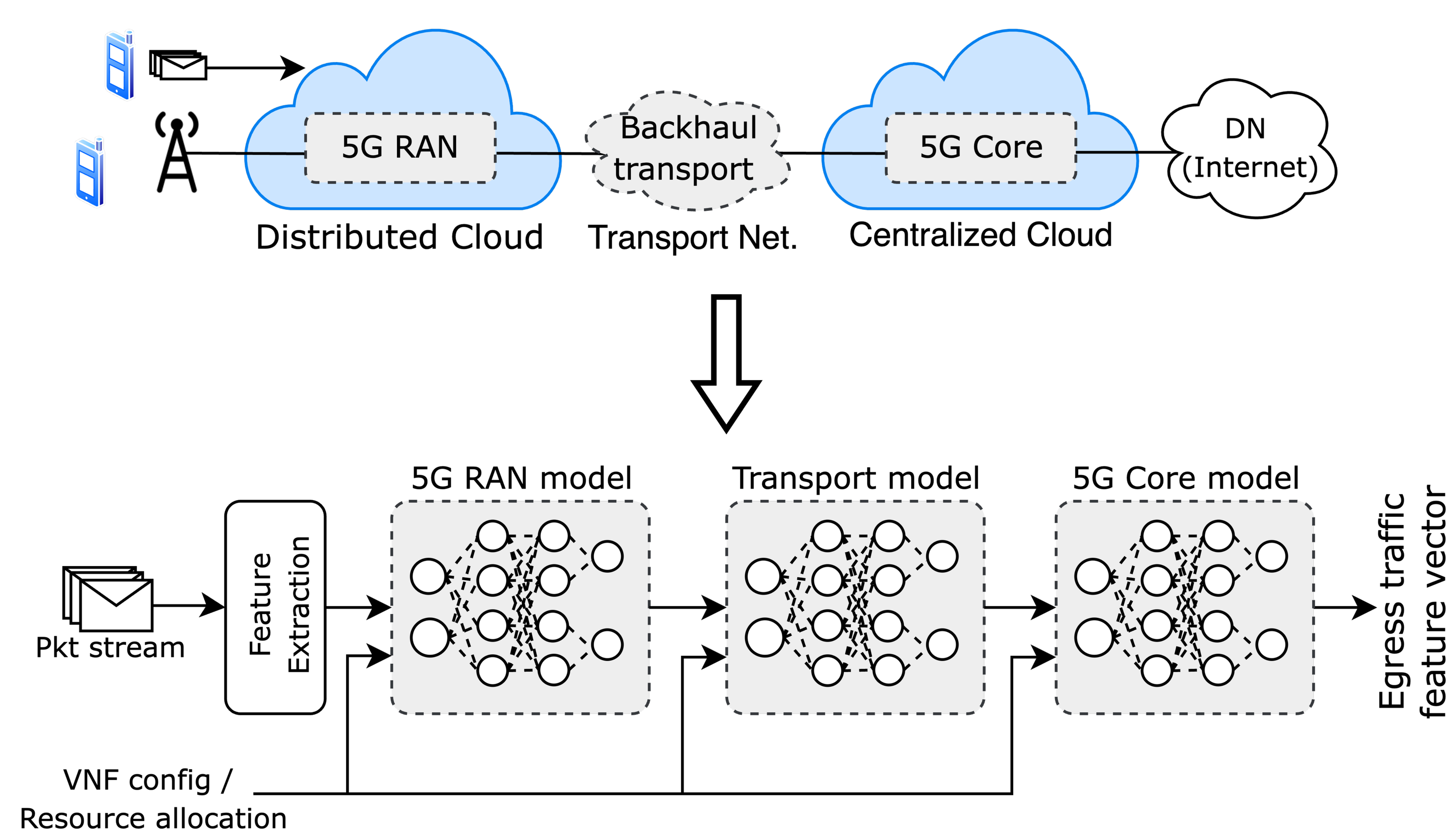
ML-BASED VNF & SLICE MODELING

VNF MODELING

- **Input/Output:** Predicts the egress traffic feature vector given input traffic trace (i.e., pcap) and VNF configuration & resource allocation
- **Dataset:** Uses data from in-lab 5G testbed and partner network operator's real network for training VNF models
- **Differentiability:** Leverages reparameterization trick to maintain VNF model differentiability
- **Composability:** Allows stacking VNF models to form end-to-end slice model

SLICE MODELING

- **Construction:** Composed using individual VNF models



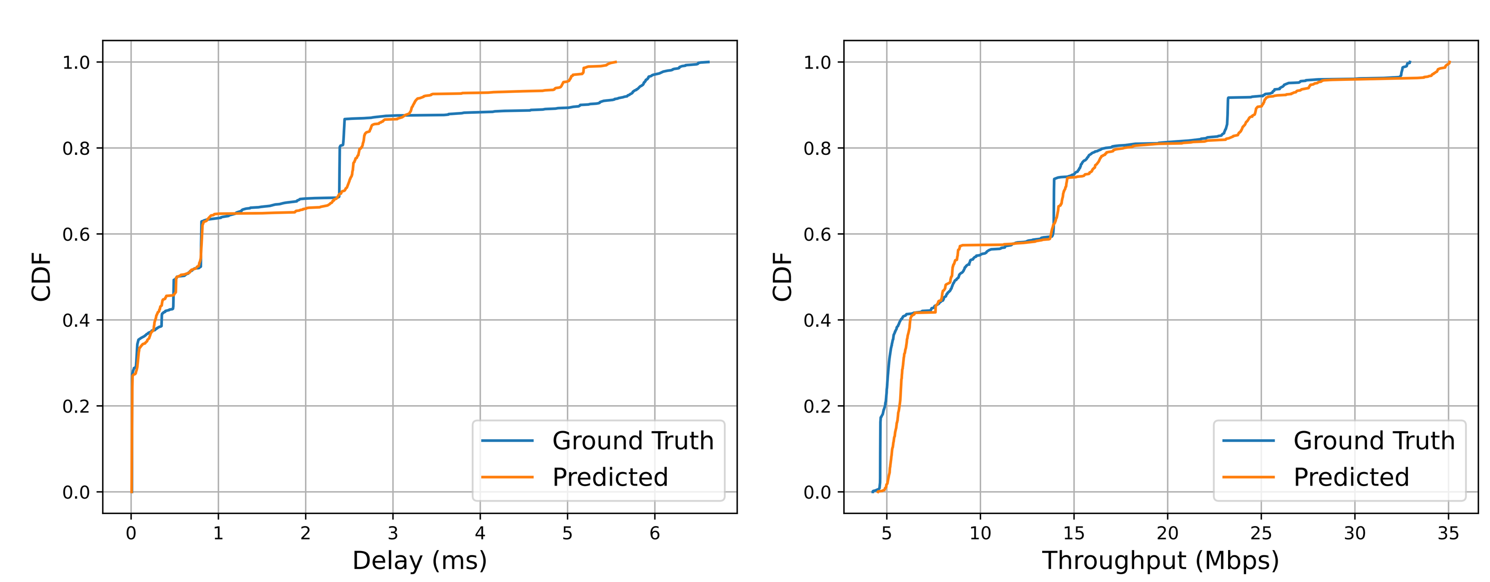


Fig: Ground truth vs. predicted throughput and delay for end-to-end slice model

RESOURCE ALLOCATION ALGORITHM

ALGORITHM

- **Lagrangian primal dual algorithm coupled with Gradient descent**
  - **Relaxed Lagrangian formulation:**

    $$\mathcal{L} = \boldsymbol{r} + \boldsymbol{\lambda} \cdot (QoS_{required} - QoS_{predicted})$$

  - **Outer loop:** Updates Lagrangian variables ($\boldsymbol{\lambda}$) based on QoS constraint violation (i.e., $QoS_{required} - QoS_{predicted}$)
  - **Inner loop:** Utilizes gradient descent to minimize the relaxed Lagrangian, with gradients from the differentiable slice model

Algorithm 1 MicroOpt Algorithm

Input: Traffic $\boldsymbol{x}_i^s$, Slice Model $f_{QoS}^s(\boldsymbol{x}_i^s, \boldsymbol{r}_i^s)$, QoS threshold q_{thresh}^s, QoS degradation threshold β_{thresh}^s, $\tau_{1,max}, \tau_{2,max}, \alpha_1, \alpha_2, \alpha_3, \epsilon_1, \epsilon_2$

Output: Optimal resource allocation vector $\boldsymbol{r}_i^s$

```
Initialize λ, μ, LB = 0, UB = ∞, τ1 = 0, τ2 = 0
while (UB−LB)/UB > ε1 or τ1 < τ1,max do            ┐ Outer loop
    r ← Initialization(x_i^s, f_QoS(x_i^s, r))     │
    while |∇_r L̂| > ε2 or τ2 < τ2,max do           │ ┐ Inner loop
        r ← [r − α1 ∇_r L̂]^+                        │ │
        τ2 ← τ2 + 1                                  │ ┘
    end while                                        │
    λ_s ← [λ_s + α2(β_i^s − β_thresh^s)]^+, ∀s       │
    μ_k ← [μ_k + α3(Σ_{s∈S} r^{s,k} − R^k)]^+, ∀k    │
    LB = max(LB, L(r, μ, λ))                         │
    UB = min(UB, Σ_{s∈S} η^T r^s)                    │
    τ1 ← τ1 + 1                                      ┘
end while
return r
```

RESULTS

- **Resource saving:**
  14.60% and 20.74% improvement over previous SOTA and peak resource allocation
- **Significantly faster resource scaling:**
  2-3 orders of magnitude faster resource scaling compared to previous SOTA