Black-Box Evasion Attacks on Data-Driven Open RAN Apps: Tailored Design and Experimental Evaluation

Pranshav Gajjar*, **Molham Khoja***, Abiodun Ganiyu, Marc Juarez, Mahesh Marina, Andrew Lehane, Vijay K Shah

NC STATE UNIVERSITY

THE UNIVERSITY of EDINBURGH informatics

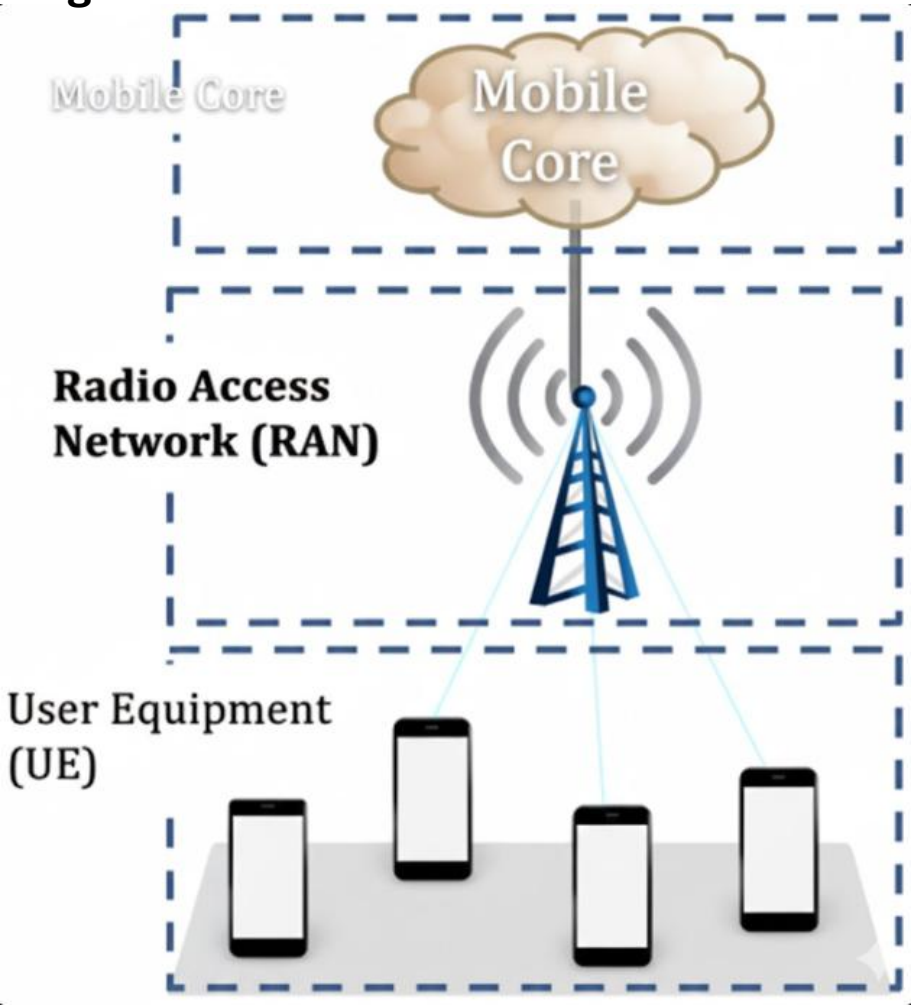
KEYSIGHT

Overview

- Explore **why the industry is adopting Open RAN** and **how its openness introduces new security challenges**.
- Present a **novel attack** designed to exploit Open RAN while meeting its **strict latency requirements**.
- Evaluate the attack using **Keysight RICtest emulator** and an **Open RAN testbed**.
- Assess the **effectiveness of common AI-based defenses** and show that they **provide only partial protection**.

Traditional RAN Makeup and Management Limitations

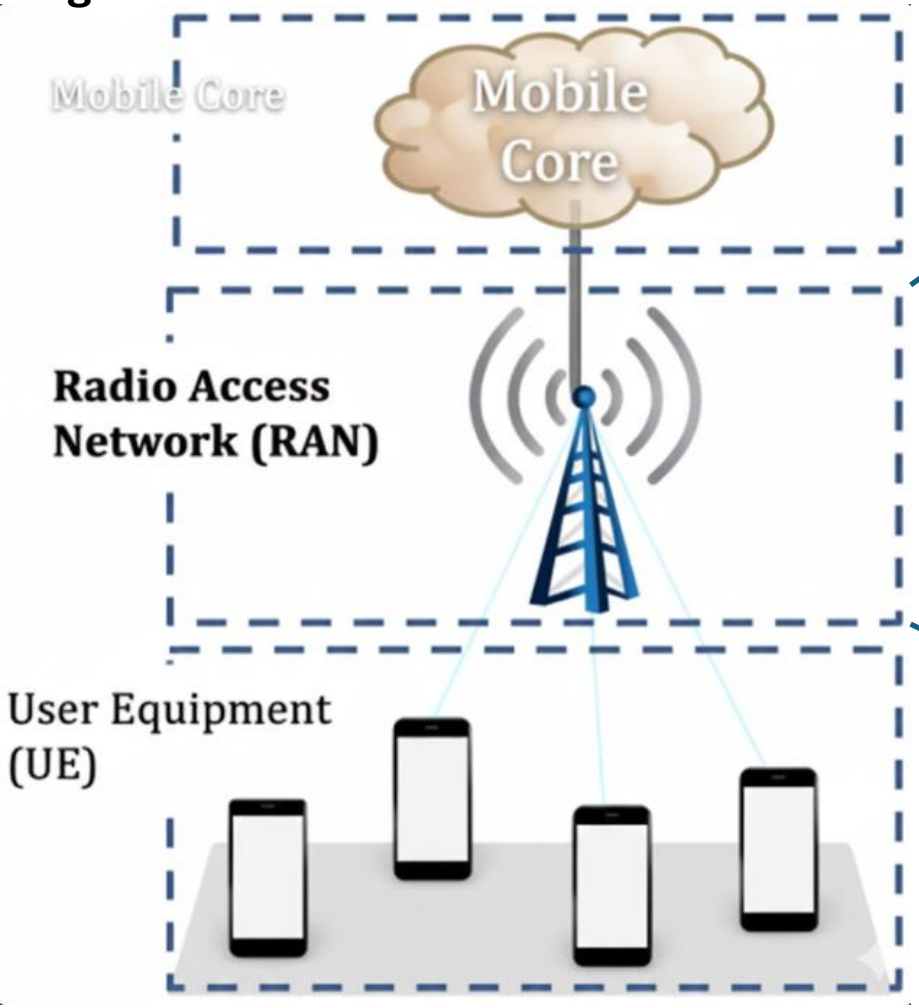
High-level view of a mobile network architecture



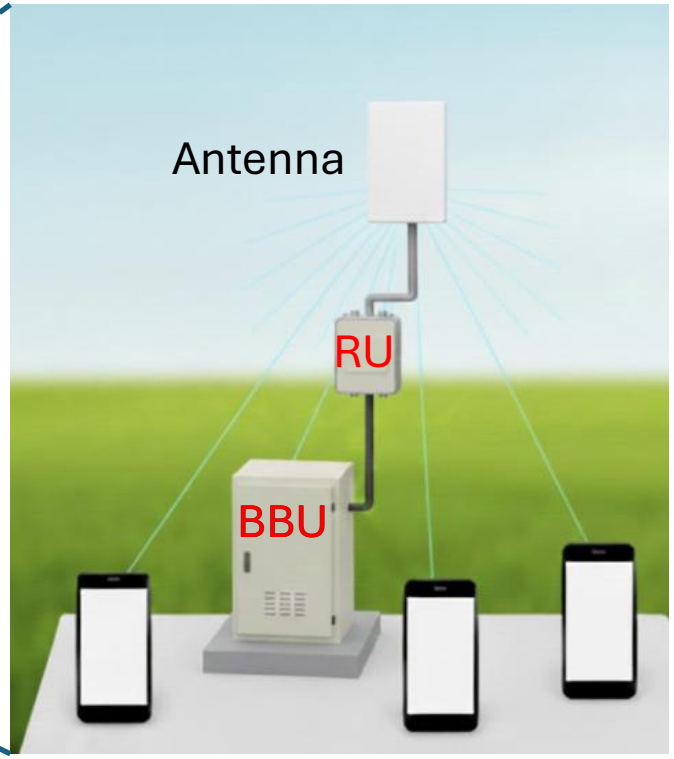
Images are AI-generated

Traditional RAN Makeup and Management Limitations

High-level view of a mobile network architecture



Traditional RAN

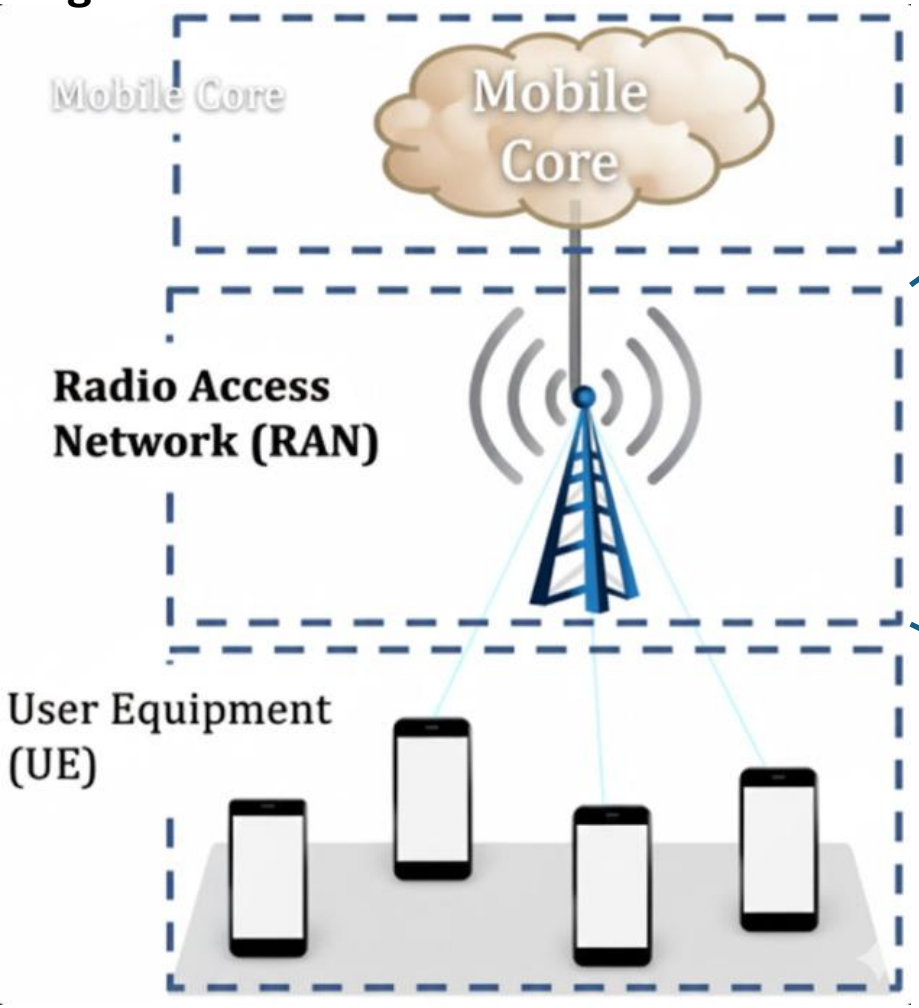


- Single-vendor lock-in → Supply chain risks
- Limited third-party automation

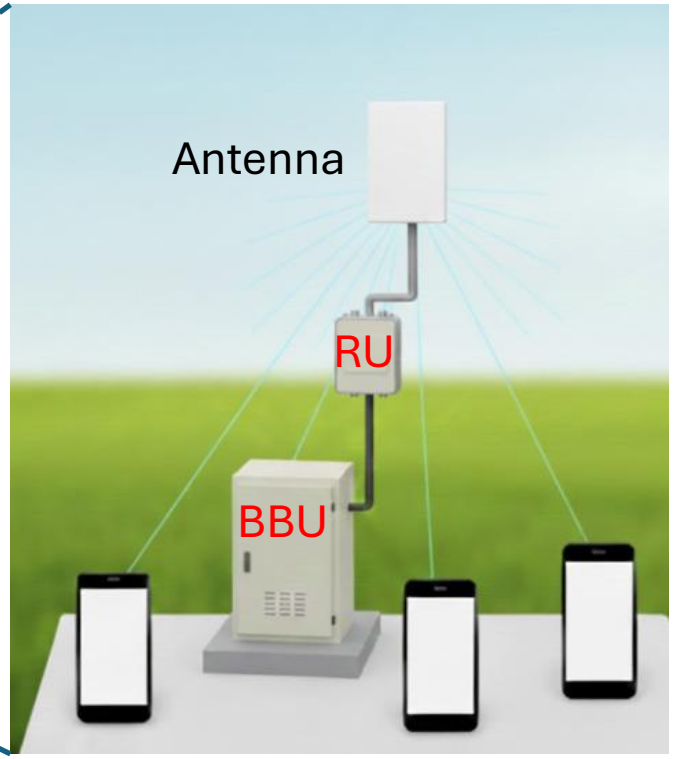
Images are AI-generated

Traditional RAN Makeup and Management Limitations

High-level view of a mobile network architecture



Traditional RAN

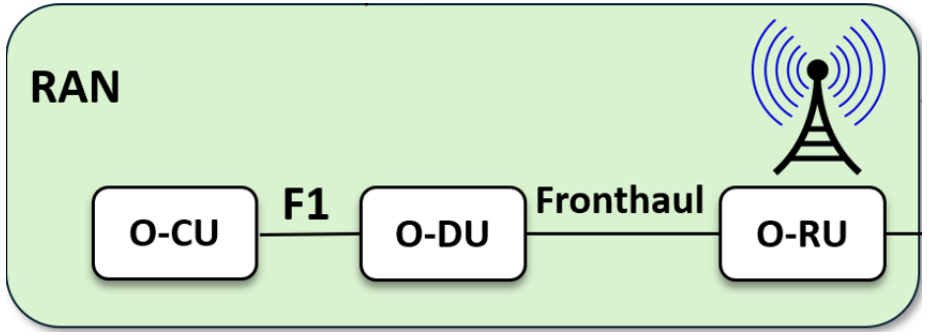


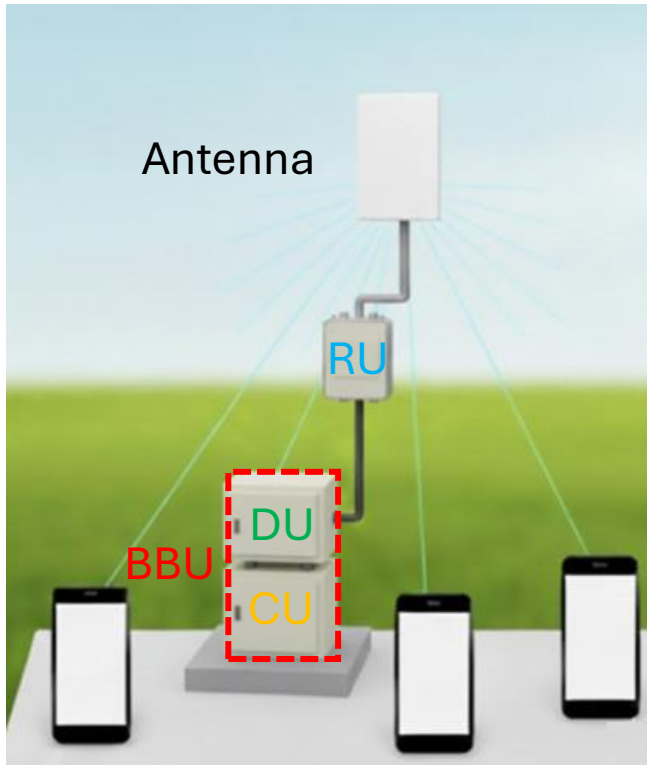
- Single-vendor lock-in → Supply chain risks
- Limited third-party automation

Fueling the shift towards "Open" RAN being standardized by O-RAN Alliance

Images are AI-generated

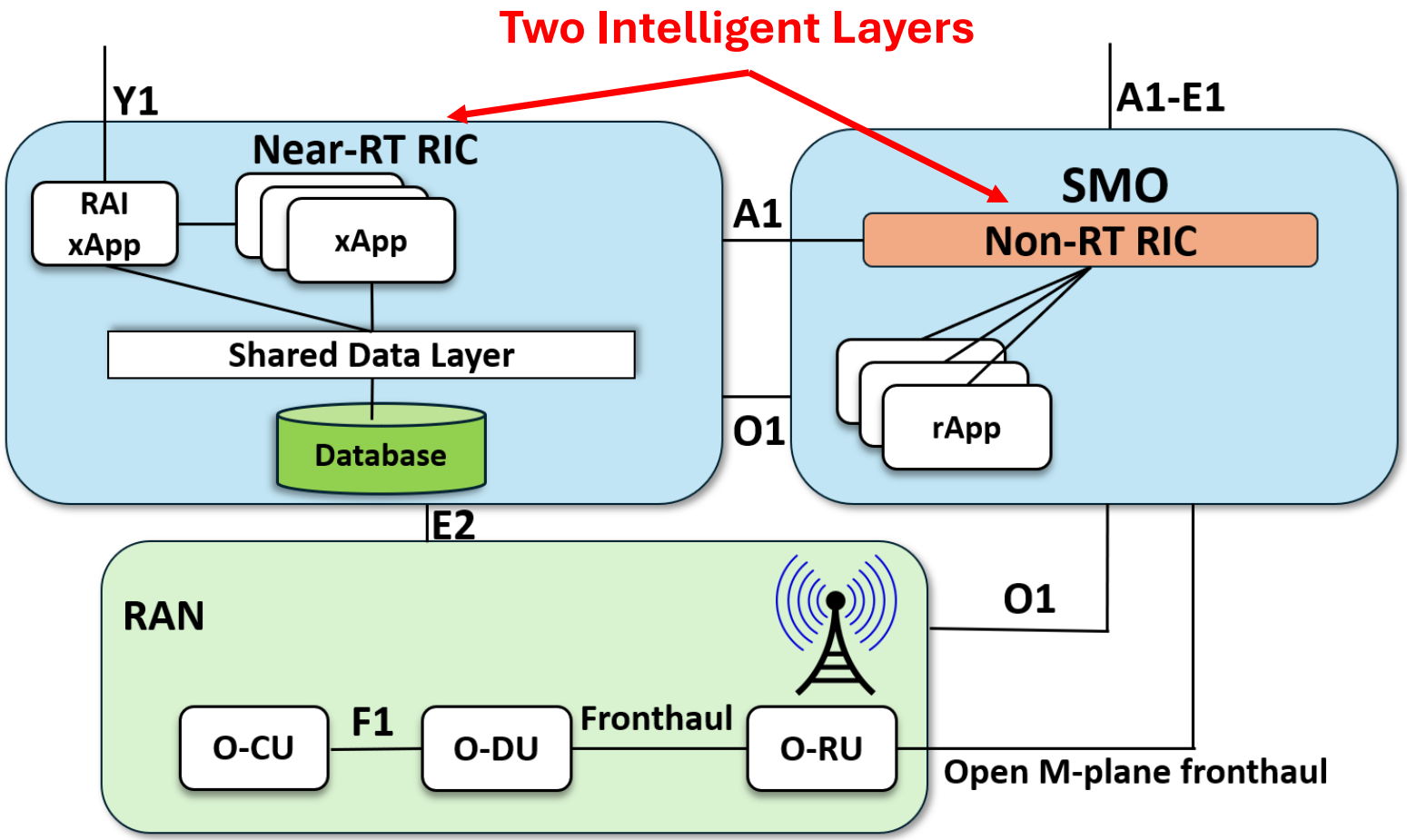
O-RAN Architecture





O-RAN alliance Introduces openness and interoperability by enabling **multi-vendor RAN components**

O-RAN Architecture



RAN Intelligent Controllers (RICs)

- Near-real-time RIC (10ms - 1s) - xApp
- Non-real-time RIC (>1s) - rApp

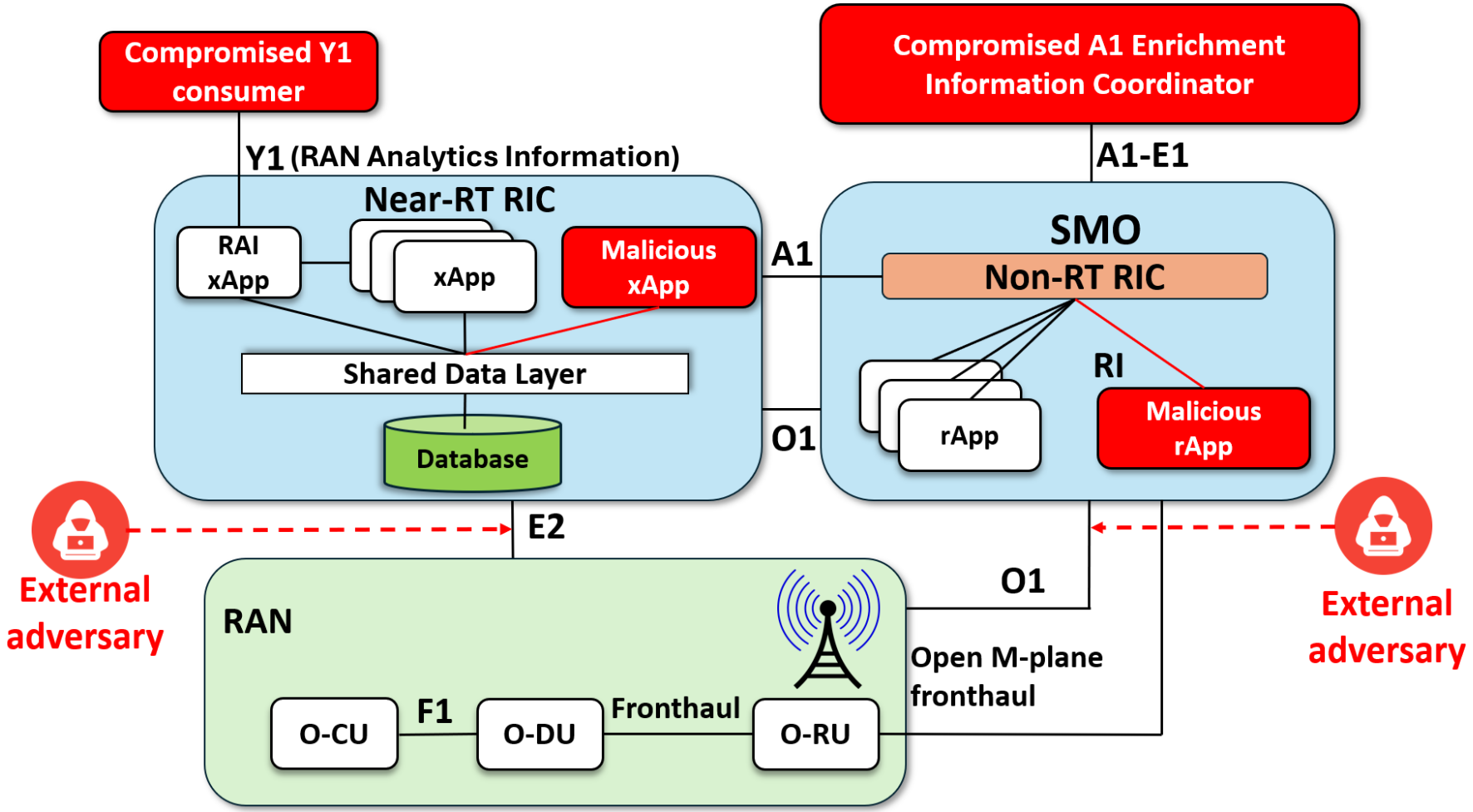
Enables intelligence and automation through RIC-driven ML (xApps/rApps).

Tracking the Global Momentum of Open RAN

Visualizing worldwide O-RAN activity across all deployment phases.

https://map.o-ran.org/

O-RAN Threat Model



- O-RAN's **open interfaces**, and **third-party app ecosystem** create **a broad attack surface**.
- In this work, we **evaluate the ML threat model** for applications in **both RICs**.

Objectives and Constraints of the Black-Box Attack

Attack Objective:

Design a **malicious app** can **attack another benign app's ML model**, causing it to produce either of the following:

- an **incorrect, random output** (a non-targeted attack)
- a **specific incorrect output** (a targeted attack)

Attack Constraints:

We designed a malicious-app attack on a benign app hosting an ML model, under these constraints:

- We **cannot access** the **model architecture or gradients**.
- We have **zero visibility** into the benign app **training dataset**.
- We are **not allowed** to **query the benign app**.

Given these constraints, any attempt to attack a benign app must **rely on a black-box approach**, as white-box methods are not possible in this setting.

Threats and Vulnerabilities in RAN and O-RAN

Capabilities of a Malicious App:

- A malicious app can **observe the input and output** of a benign app.
- A malicious app can **inject crafted noise into a benign app's input** during **inference**.

Potential Vulnerabilities Enabling Malicious Actions:

The malicious app can carry out the above actions by **exploiting potential RAN security risks**, such as:

- Misconfigured access policies
- Basic security practices are sometimes overlooked (e.g., default passwords)

Example of Potential Security Risks in RAN/O-RAN System

- Misconfigured access policies

This article is more than **5 months old**

Virgin Media O2 mobile users' locations exposed for two years in security flaw

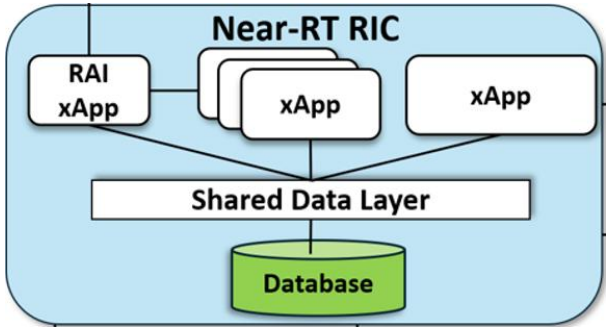
Most precise tracking was in cities, where mobile masts cover areas as small as 100 square metres

https://www.theguardian.com/media/2025/may/29/virgin-media-o2-mobile-users-locations-exposed-for-two-years-in-security-flaw

Challenges

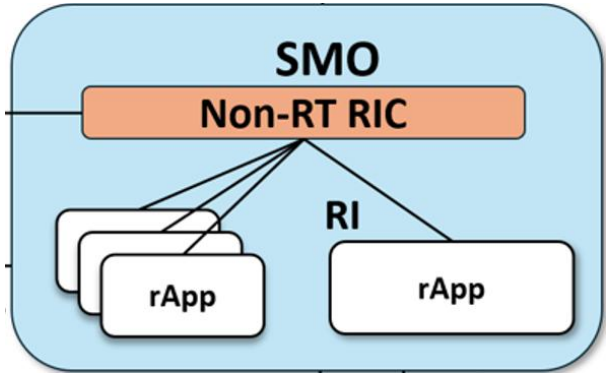
xApp Attack:

- xApps run **fast control loops** (10ms - 1s)
- Attacks must be **crafted and injected within milliseconds**



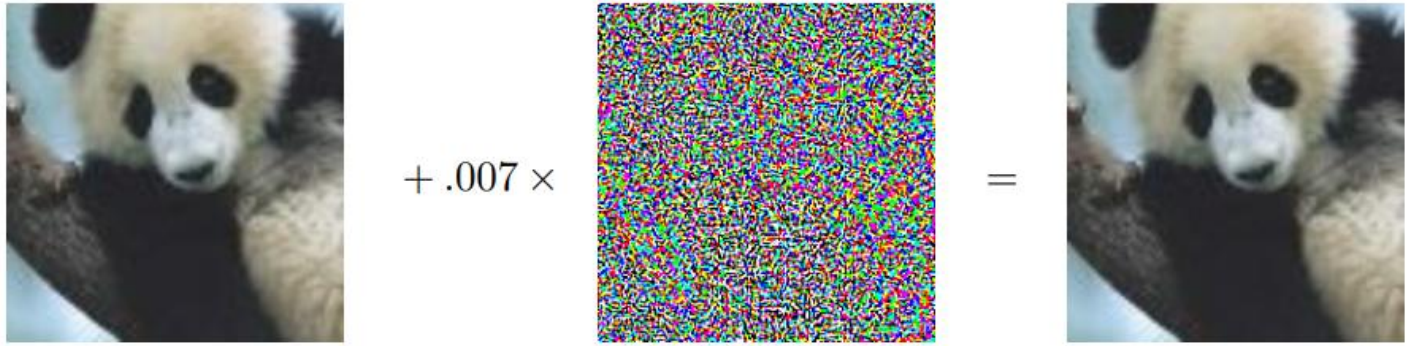
rApp Attack:

- rApps manage a **large number of cells**.
- Attacks must be crafted and injected for **precise system manipulation**, not just output changes



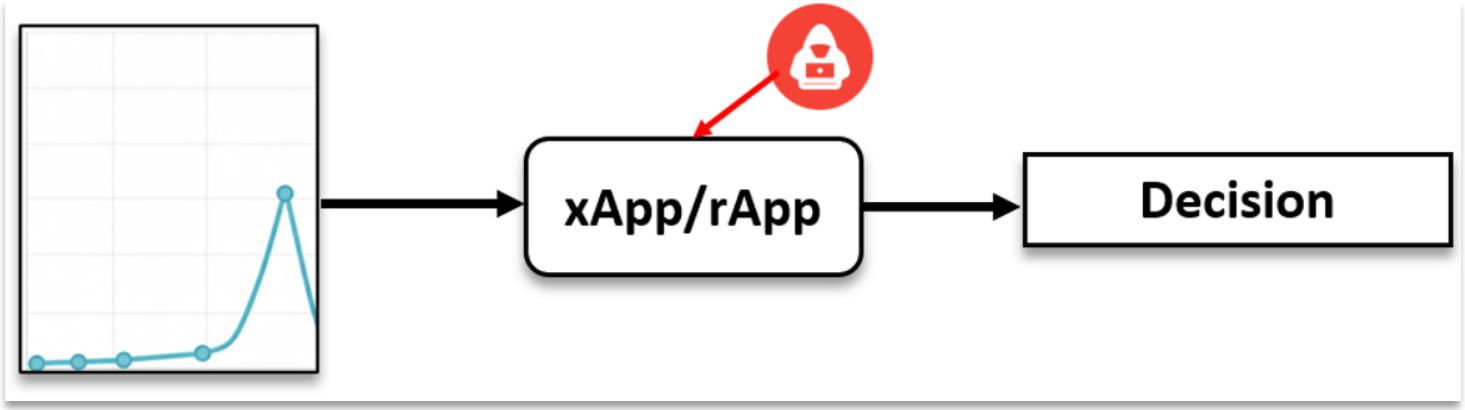
Fast Gradient Sign Method (FGSM)

The idea is to perturb the input data by adding a small amount of noise based on the gradient of the loss with respect to the input by using a technique called FGSM.



Goodfellow IJ, Shlens J, Szegedy C. Explaining and harnessing adversarial examples. arXiv preprint arXiv:1412.6572. 2014 Dec 20

Limitations of FGSM in the ORAN Context



- FGSM requires model architecture or gradients → usually not visible in O-RAN
- FGSM must generate new noise for each input → too slow for millisecond-level (10ms - 1s) control loops.

Note: Further details are provided in the Network Performance Results section of the paper.

Design of Framework for xApp/rApp Attack

1. A malicious xApp/rApp passively monitors the target benign app and **collects input–output pairs** for training data.

Design of Framework for xApp/rApp Attack

1. A malicious xApp/rApp passively monitors the target benign app and **collects input–output pairs** for training data.

2. The **Model Cloning Algorithm trains surrogate models** (e.g., ResNet, MobileNet) using the **collected dataset**.

Note: Detailed explanations of the Model Cloning Algorithm is provided in the paper.

Design of Framework for xApp/rApp Attack

1. A malicious xApp/rApp passively monitors the target benign app and **collects input–output pairs** for training data.

2. The **Model Cloning Algorithm trains surrogate models** (e.g., ResNet, MobileNet) using the **collected dataset**.

3. The malicious app then **selects one of the two algorithms to generate a perturbation mask** based on the attack type.

    - **Universal Adversarial Perturbation (UAP):** for non-targeted attacks.

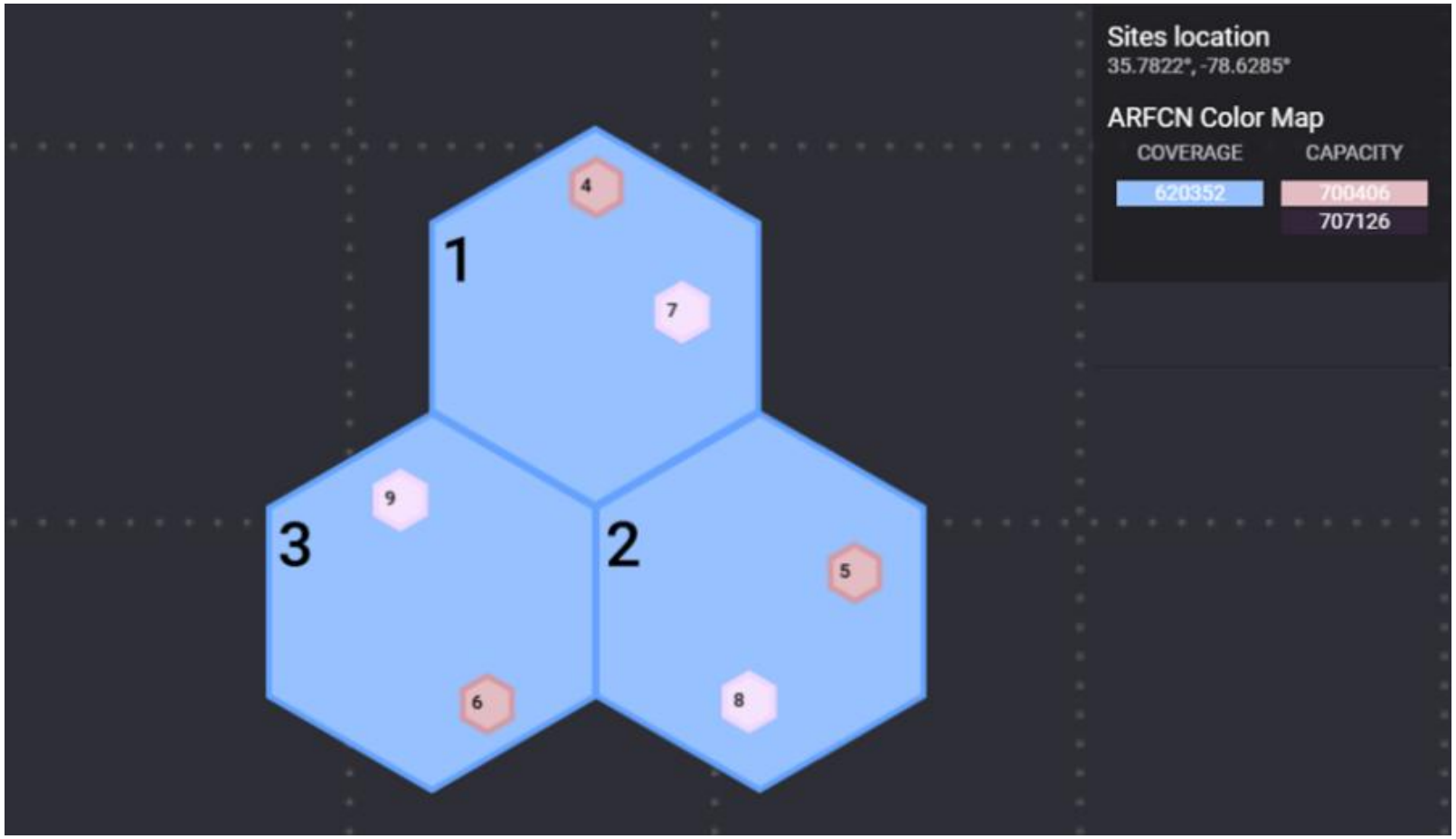
    - **Targeted Universal Perturbation (TUP):** for targeted attacks.

Note: The **universal perturbation**, once generated, can be **applied to all KPI inputs** for a given xApp/rApp **without needing to be recreated**. This universal perturbation pushes the model to make **incorrect decisions**

Note: The paper provides detailed explanations of the UAP and TUP algorithms, along with the proposed framework.

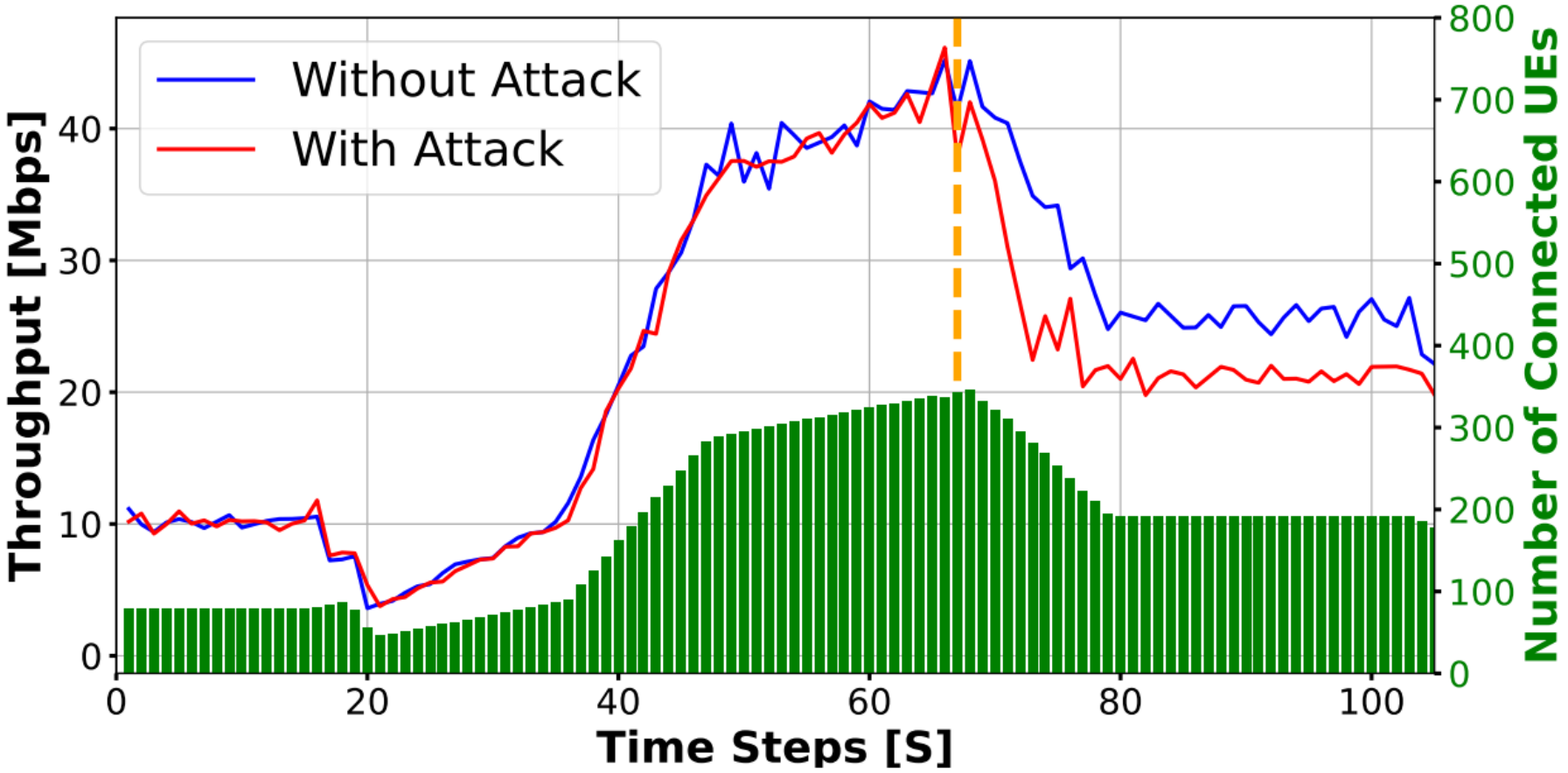
Evaluation

Keysight RICtest Emulator



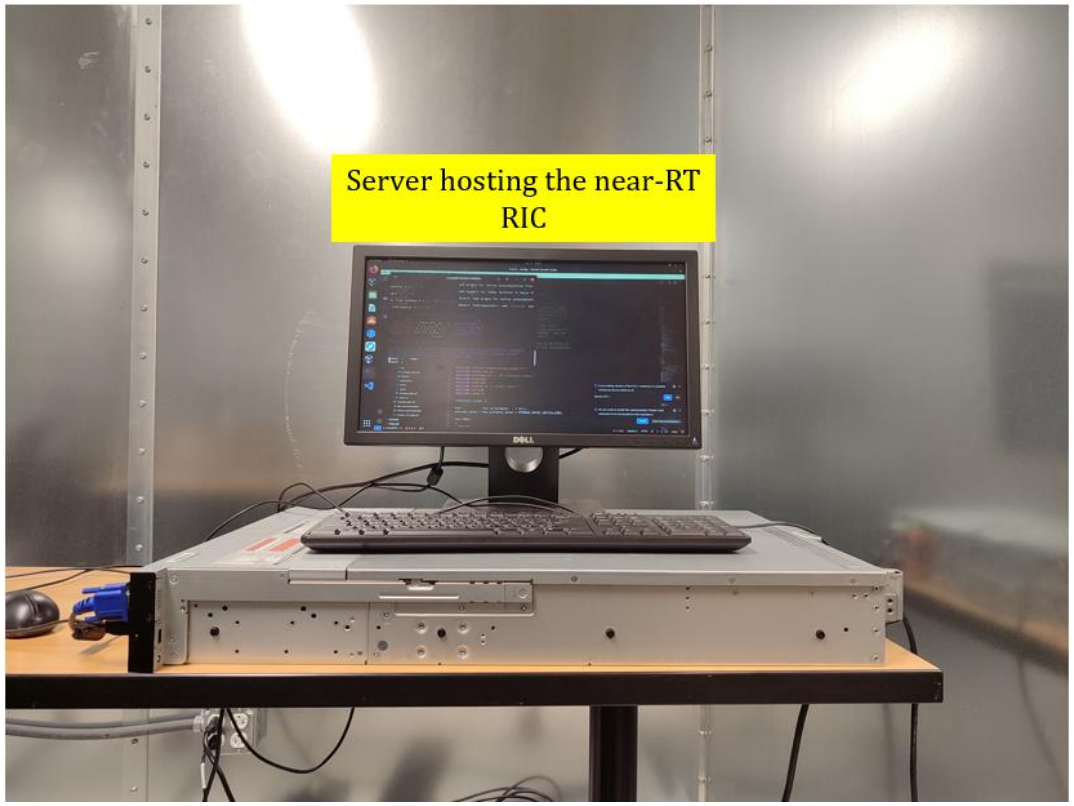
- **UE Distribution:** 10 UEs per coverage cell; 0–55 UEs per capacity cell
- **Traffic Pattern:** Mainly downlink; mix of steady and bell-curve patterns.
- **Power-Saving rApp:** Turns capacity cells on/off based on cell load.

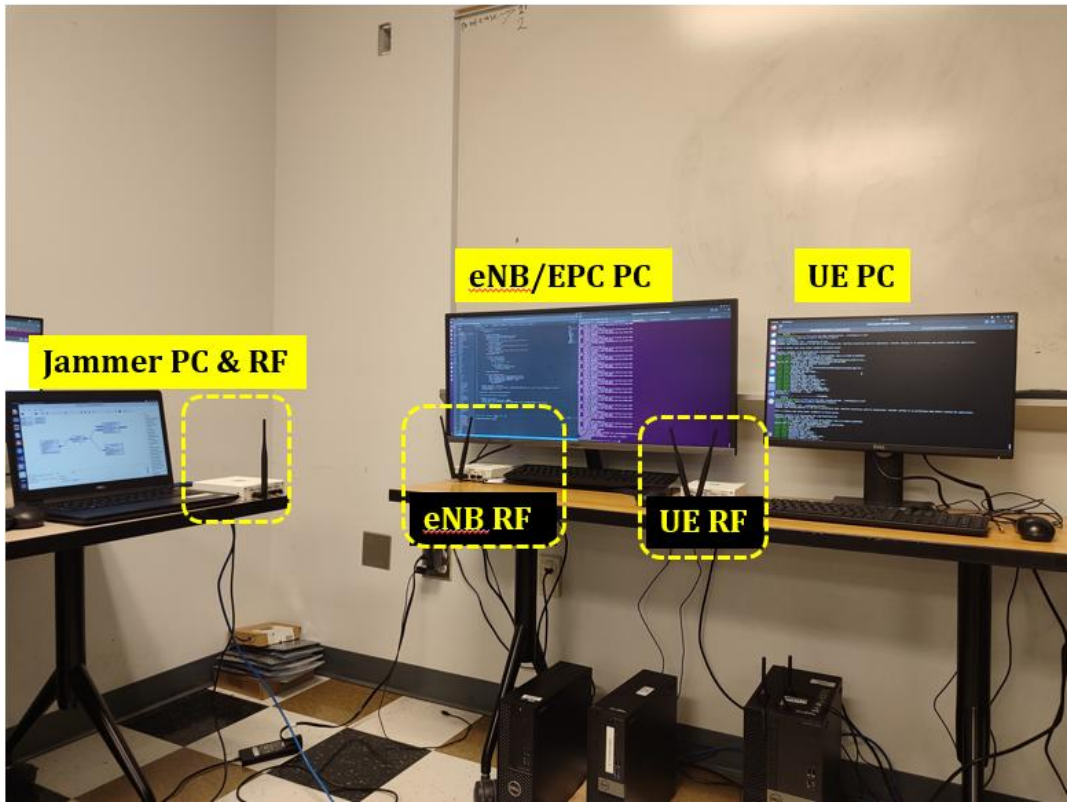
Performance Degradation Caused by Attacks on the Power-Saving rApp



- **Normal Operation:** Both coverage and capacity cells active → DL throughput varies with user load.
- **Under Attack :** Attacker deactivates 2 capacity cells → users shift to coverage cells → coverage cell overloaded → significant drop in DL throughput.

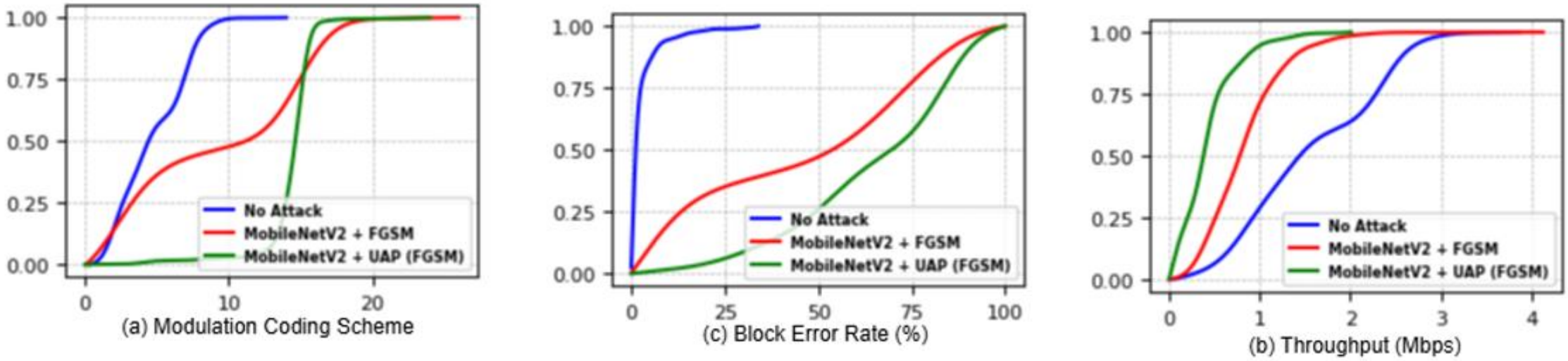
LTE/5G O-RAN Testbed at NC State University





- UE connected to the RAN, and uplink traffic was generated using iperf3
- Jammer is activated
- The Interference Classification (IC) xApp is activated to detect uplink interference caused by jammer

Performance Degradation Caused by Attacks on the Interference Classification xApp



- **Normal Operation:** IC xApp detects interference → RAN uses adaptive MCS → maintains low error rates & optimal throughput.
- **Under Attack:** xApp fails to detect interference → RAN uses fixed MCS → higher BLER & lower throughput.
- **UAP vs. FGSM:** UAP is more effective at targeting xApps; FGSM is slower, allowing xApp to sometimes make correct predictions.

Evaluation of Defense Mechanisms

Defense Methods Evaluated:

- Adversarial Training (AT)

- Defensive Distillation.

Results:

Our black-box evasion strategy still **bypasses these defenses**, but it needs a **larger perturbation** to do so

Summary

- **Developed** an effective **black-box evasion attack** targeting ML-based Apps in near-RT and non-RT RICs**.**

- **Evaluated** the attack using the **Keysight emulator** and the **O-RAN testbed**.

- Demonstrated **strong attack success even against key defenses methods** such as **defensive distillation** and **adversarial training.**

Happy to take questions