Towards Data Driven and AI Powered Open RAN

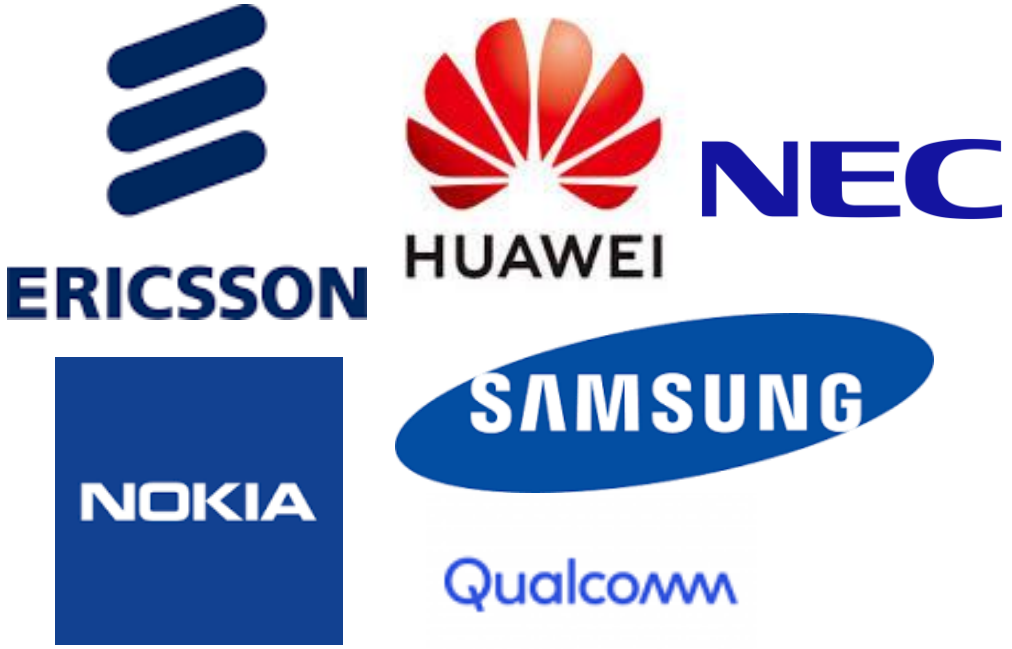
Mahesh K. Marina

Joint work with Chuanhao Sun, Ujjwal Pawar, et al.

THE UNIVERSITY of EDINBURGH
informatics

icsa | Institute for Computing Systems Architecture

Traditional Vendor Ecosystem



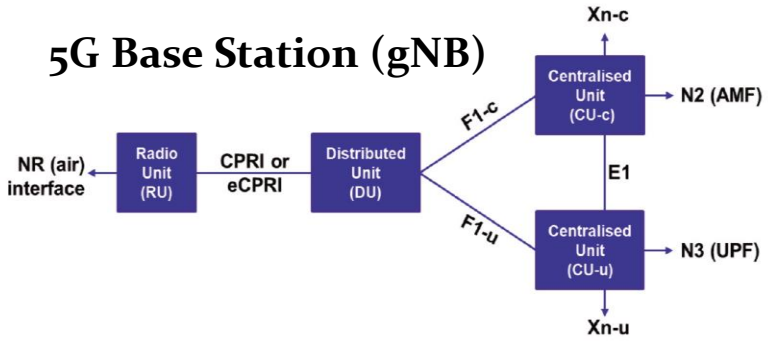
Emerging Vendor Ecosystem

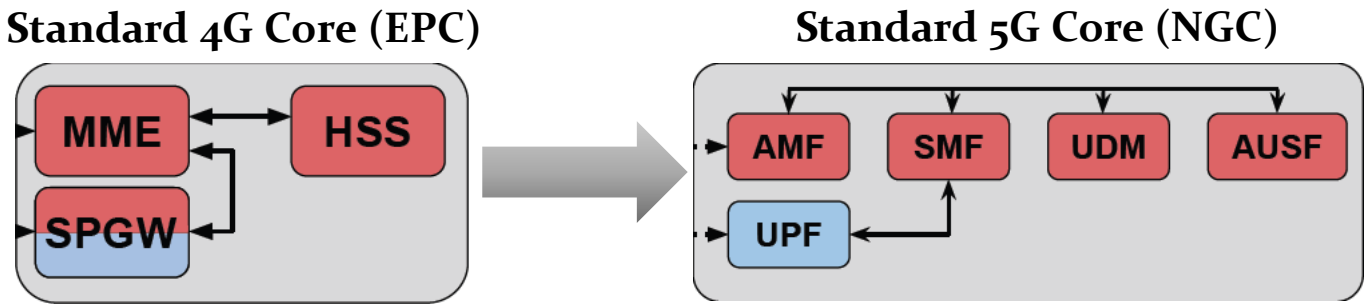
Economics the key driver for change:
- cost reduction
- new value addition & revenue generation

“Disaggregation” – the key enabler

Three Levels of Disaggregation

1. **Disaggregate network function software from hardware**
   - Run network function software on cheaper commodity hardware *aka* **Network Functions Virtualization (NFV)**
   - Deploy on (edge) **cloud** to reduce costs further
2. **Disaggregate control plane from data plane *a la* Software-Defined Networking (SDN)**
   - Flexible (programmable) network control, performant data plane
3. **Disaggregate network functions themselves → microservices**
   - Modularity, flexible scaling, easier evolution





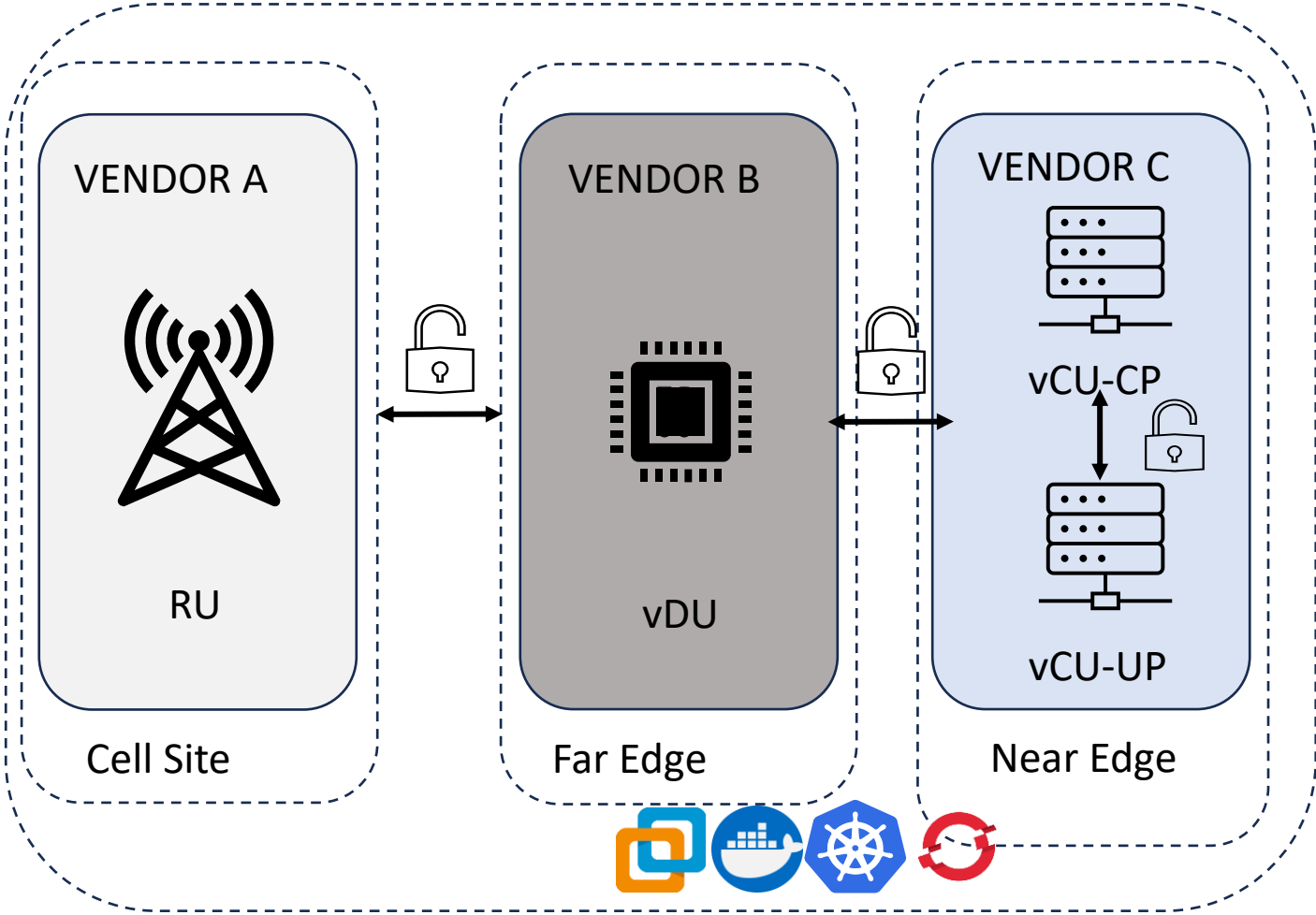
Disaggregation introduced new research challenges and opportunities

Reflected in my own series of works…

- **How to enable programmable RAN control?**
  - **FlexRAN** software-defined RAN platform **[CoNEXT'16]**
  - Open source, used by 150+ groups, forerunner to operator-driven O-RAN architecture
- **How to virtualize radio resources?**
  - **Orion** RAN hypervisor for radio resource virtualization and RAN slicing **[MobiCom'17]**
- **How to realize intelligent and adaptive resource control applications over the above capabilities?**
  - **Iris** indoor neutral-host small cell spectrum sharing system, leveraging Orion and FlexRAN systems **[JSAC'19]**
- **How to enable cost-effective and high-capacity wireless backhaul?**
  - **WhiteHaul** software-driven white space spectrum aggregation system with commodity radio hardware **[MobiSys'20]**
- **Can we leverage the cloud for cost-effective digital twins for mobile networks?**
  - **Nervion** cloud-native RAN emulator for scalable and flexible mobile core evaluation **[MobiCom'21]**
- **How to design an efficient and scalable virtualized mobile core?**
  - **CoreKube** message focused and cloud-native mobile core system **[MobiCom'23]**

Open RAN

- Features all three levels of disaggregation
- Can be realized with multivendor components → open (diverse) ecosystem
- Got traction in the industry and becoming a reality

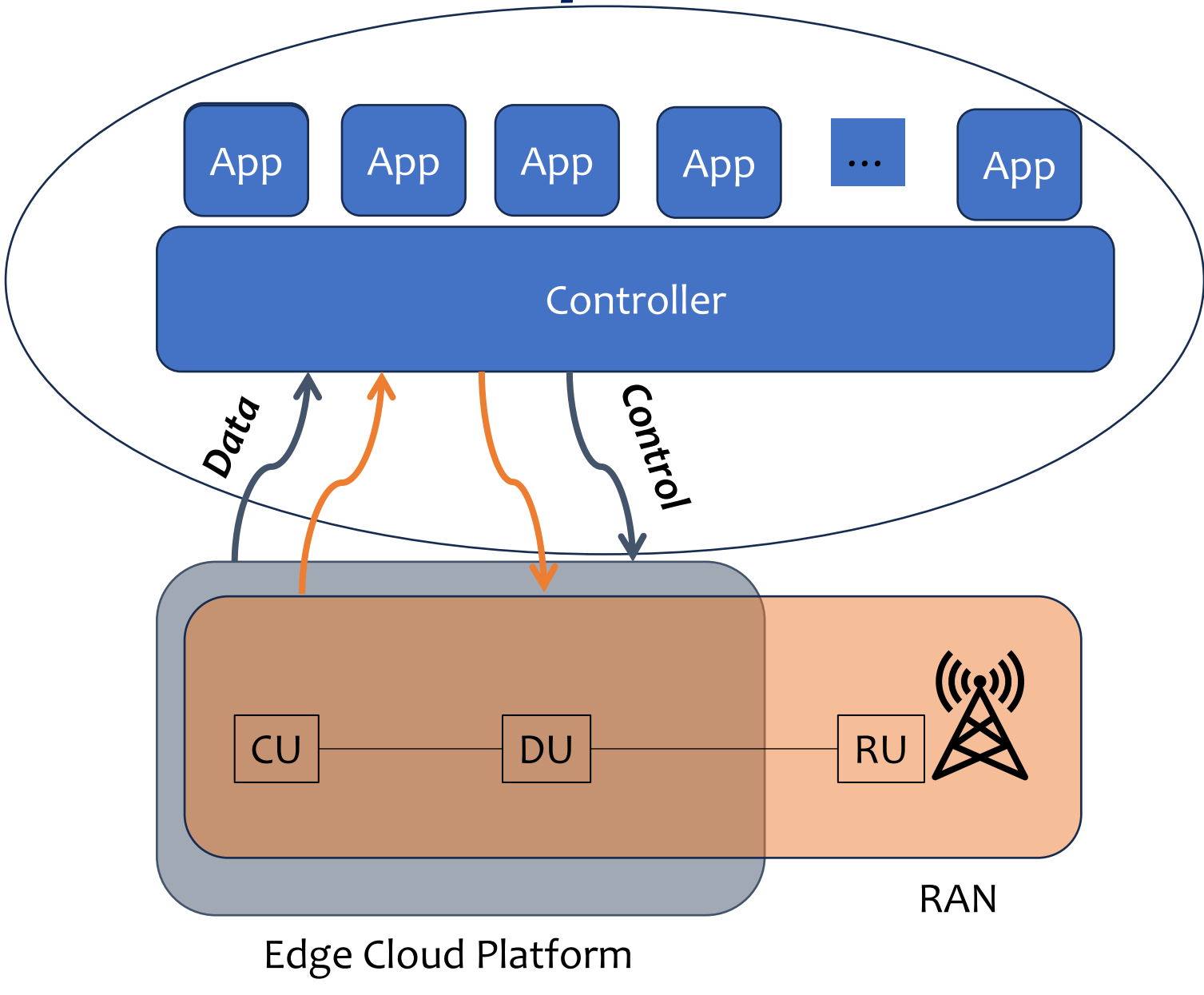


Another transformation in the RAN underway
→ data-driven and AI-powered operation

Data Driven & AI Powered Open RAN

- Rearchitect RAN control and management to be more disaggregated and App oriented
  - Drive Apps/Controller with RAN & Platform Data
  - Leverage the power of AI in Apps/Controller

- Potential for even more openness and significantly expand the ecosystem



Motivation for Data Driven RAN Operation

- **Cost reduction**
  - Automating network management
  - Optimizing resource allocation and enhancing QoE
  - Enabling energy efficient operation
  - Achieving efficient spectrum use
  - …

- **Value addition**
  - Data analytics
  - Positioning
  - Sensing
  - …

Challenges

- Data acquisition seamlessly and efficiently to enable data-driven operation
- Troubleshooting harder with more disaggregation and vendors
- Robustness of AI models to attacks
- Explainability
- Uncertainty quantification
- Generalization
- App architecture
- Edge infra

Generative AI based Solutions

Generative AI in a Nutshell

Let X represents the space of all true data points x.

Data can be of any modality – numerical, text, images, video, …

Learn the distribution $\boldsymbol{P_{model}(x)}$ *via a generative model that estimates the true but unknown data distribution* $\boldsymbol{P_{data}(x)}$*.*

Several different generative modeling approaches exist – GANs, VAEs, transformers, diffusion models, …

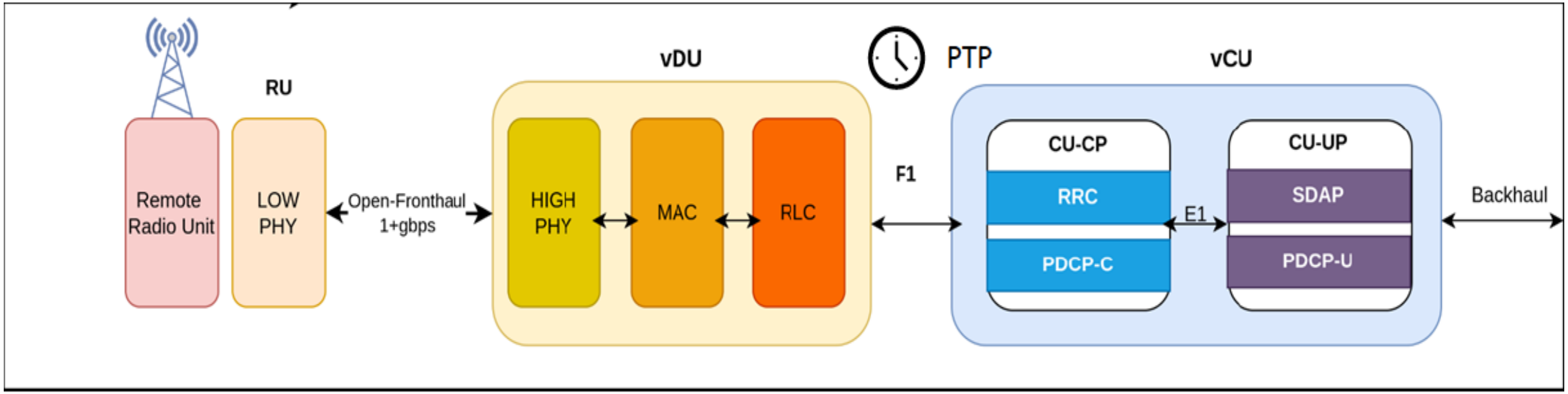
"Sample" from P_{model} *to generate/create new data points x' that mimic* $x \in X$*, i.e., appear to be drawn from* $P_{data}(x)$*.*

Generative AI for Cellular Networks

- Synthetic cellular network traffic generation:
  - **SpectraGAN** [CoNEXT'21] for spatiotemporal traffic generation
  - **CartaGenie** [PerCom'22] for traffic snapshot generation
  - **AppShot** [TNSM'22] for service-level traffic snapshot generation
- Cost-effective drive test measurement data w/ **GenDT** [CoNEXT'22]
- Open RAN anomaly detection w/ **SpotLight** [MobiCom'24]
- Efficient Open RAN telemetry w/ **NetGSR** [CoNEXT'24]

Open RAN Anomaly Detection Problem

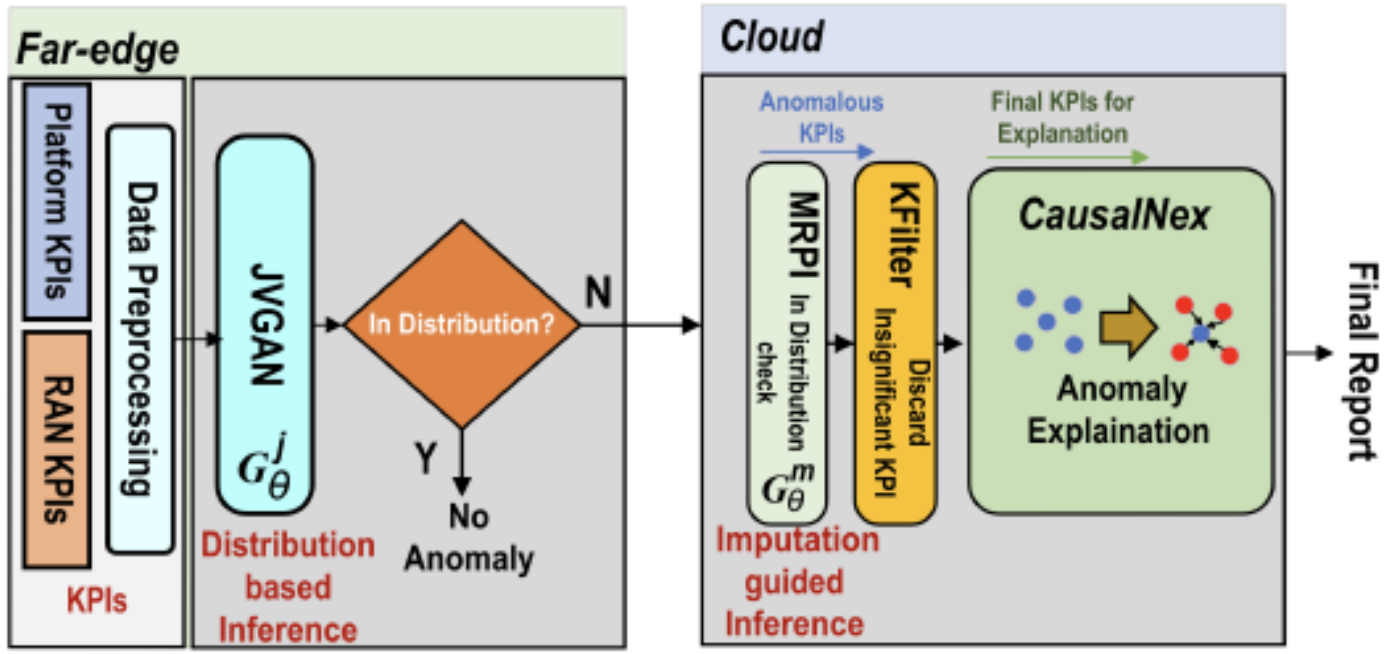
- When operational problems (anomalies) occur, how to reliably detect and pinpoint the root cause?
- In other words, which vendor's component should be blamed?



Questions to tackle…

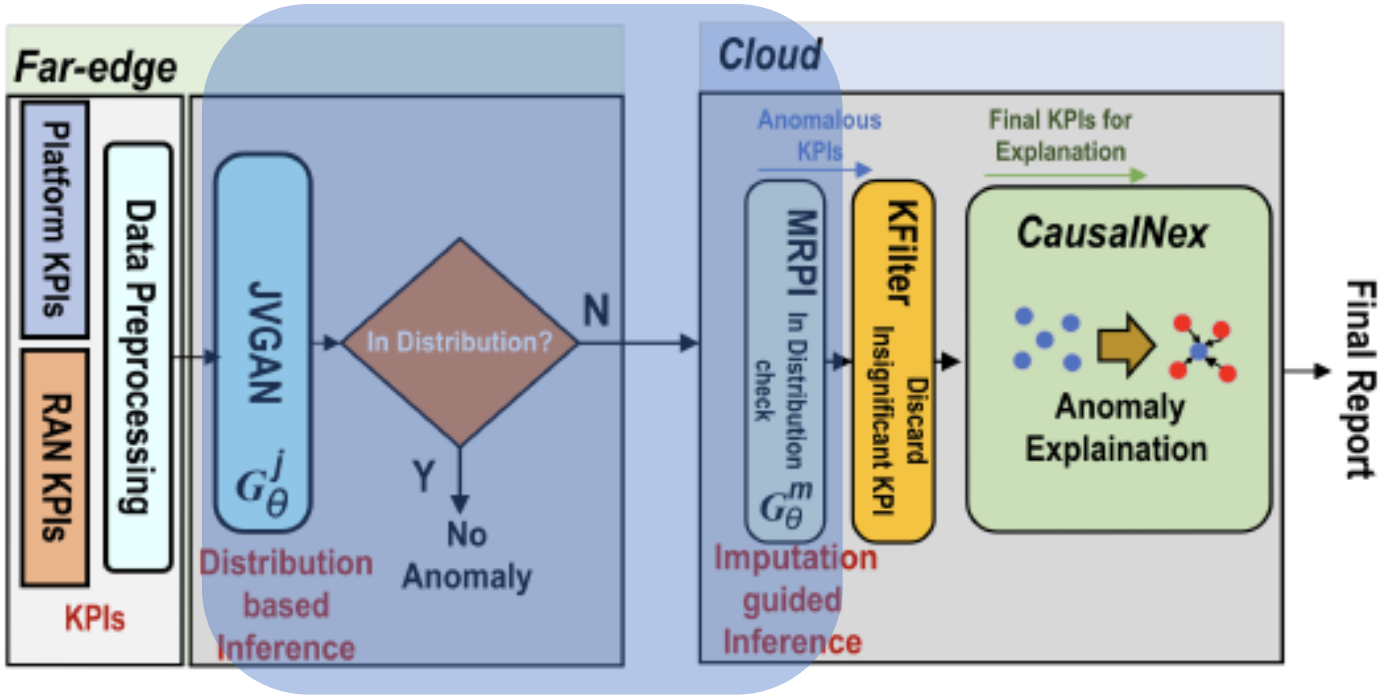
1. **Data Collection:** What data is needed for anomaly detection in Open RAN?
2. **Accuracy:** How to perform reliable anomaly detection over Open RAN data with minimal false alarms?
3. **Explainability:** When an anomaly occurs, how to ensure that root cause is identified?
4. **Efficiency:** How to make the anomaly detection system deployable by complying with limited compute resources at far edge and limiting bandwidth consumption?

Our Solution: SpotLight [MobiCom'24]



- *Holistic data collection process to get relevant KPIs from RAN and Platform*
- *New tailored GenAI based 2-stage anomaly detection method*
- *Causal discovery based root cause identification*
- *Distributed pipeline across far-edge and cloud*

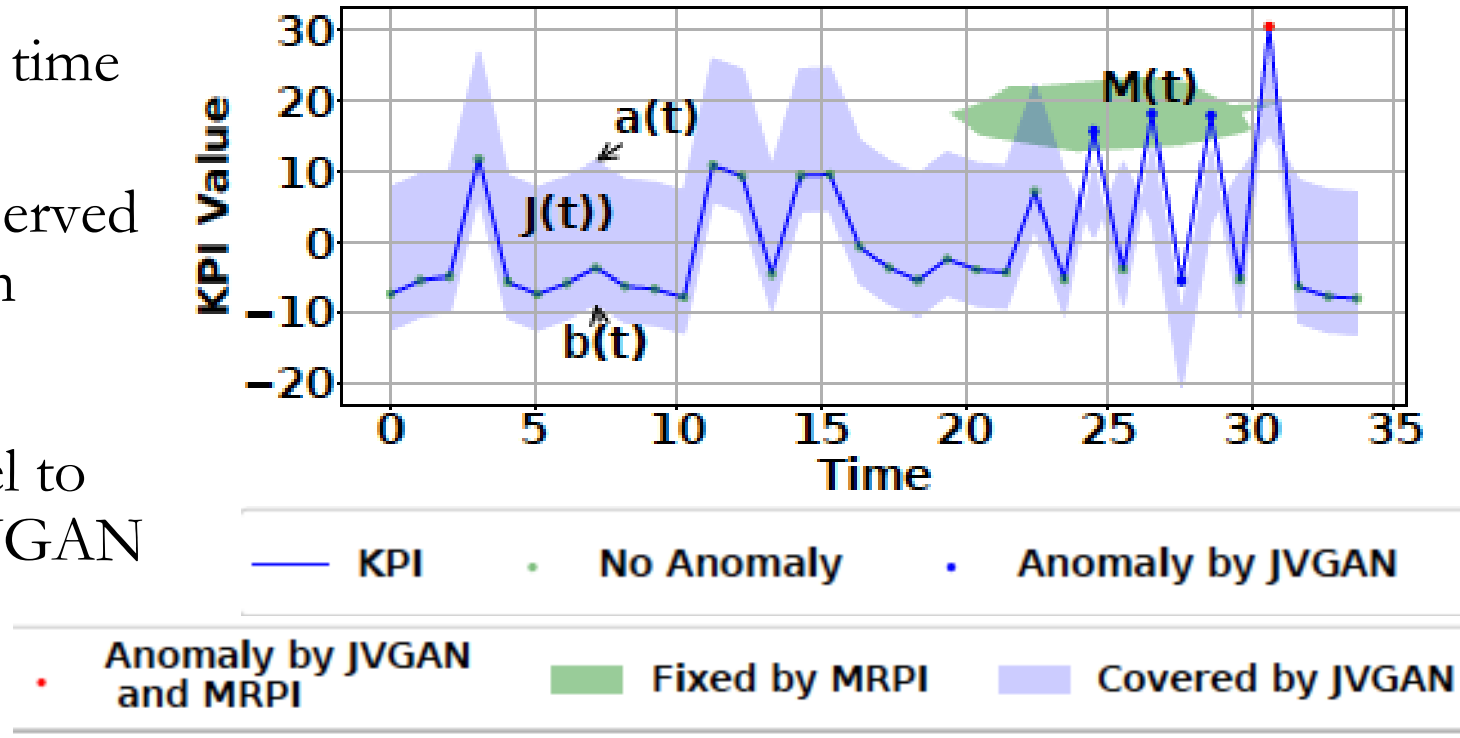
Our Solution: SpotLight [MobiCom'24]



- *Holistic data collection process to get relevant KPIs from RAN and Platform*
- ***New tailored GenAI based 2-stage anomaly detection method***
- *Causal discovery based root cause identification*
- *Distributed pipeline across far-edge and cloud*

SpotLight Anomaly Detection Illustration

- **JVGAN**
  - Learn distribution of normal KPI time series
  - **Detect potential anomaly** if observed time series not fully in distribution
- **MRPI**
  - Diffusion based imputation model to **confirm anomaly** detected by JVGAN



Generative AI for Cellular Networks

- Synthetic cellular network traffic generation:
  - **SpectraGAN** [CoNEXT'21] for spatiotemporal traffic generation
  - **CartaGenie** [PerCom'22] for traffic snapshot generation
  - **AppShot** [TNSM'22] for service-level traffic snapshot generation
- Cost-effective drive test measurement data w/ **GenDT** [CoNEXT'22]
- Open RAN anomaly detection w/ **SpotLight** [MobiCom'24]
- Efficient Open RAN telemetry w/ **NetGSR** [CoNEXT'24]

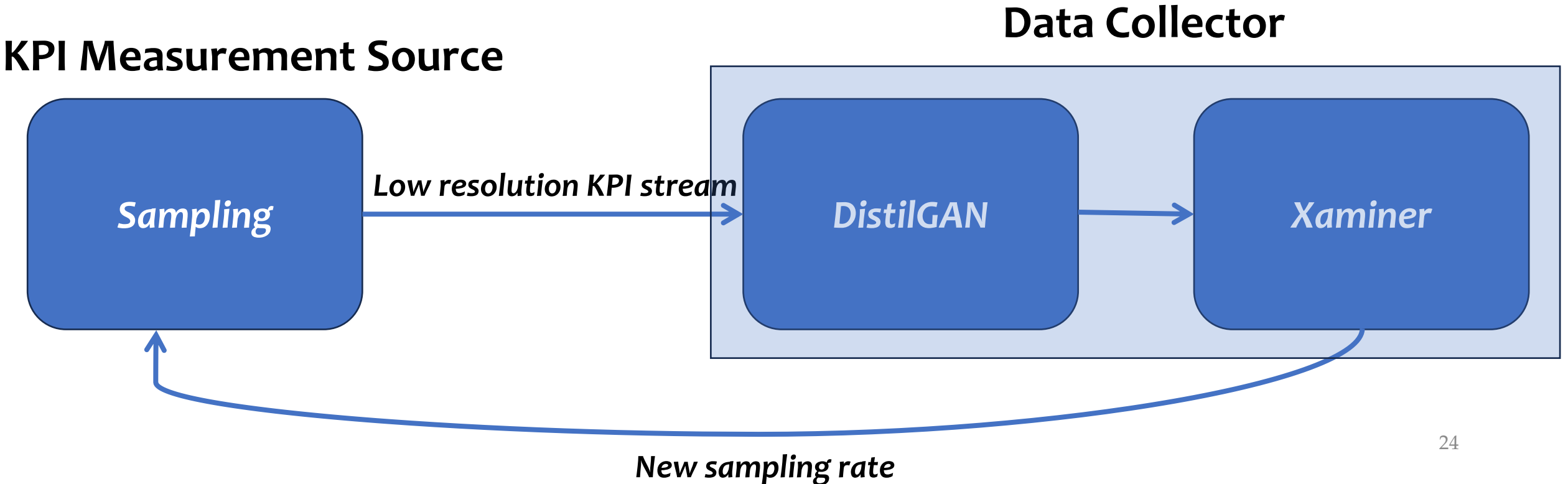
Motivation

- Potential efficiency and automation gains from data-driven operation in Open RAN hinges on scope and granularity of KPIs – "the data"
- Desired granularity KPI and App dependent but generally finer the better
- Fine-grained KPI data collection causes high overhead and can disrupt RAN operation
- Capability to measure any KPI necessary but insufficient

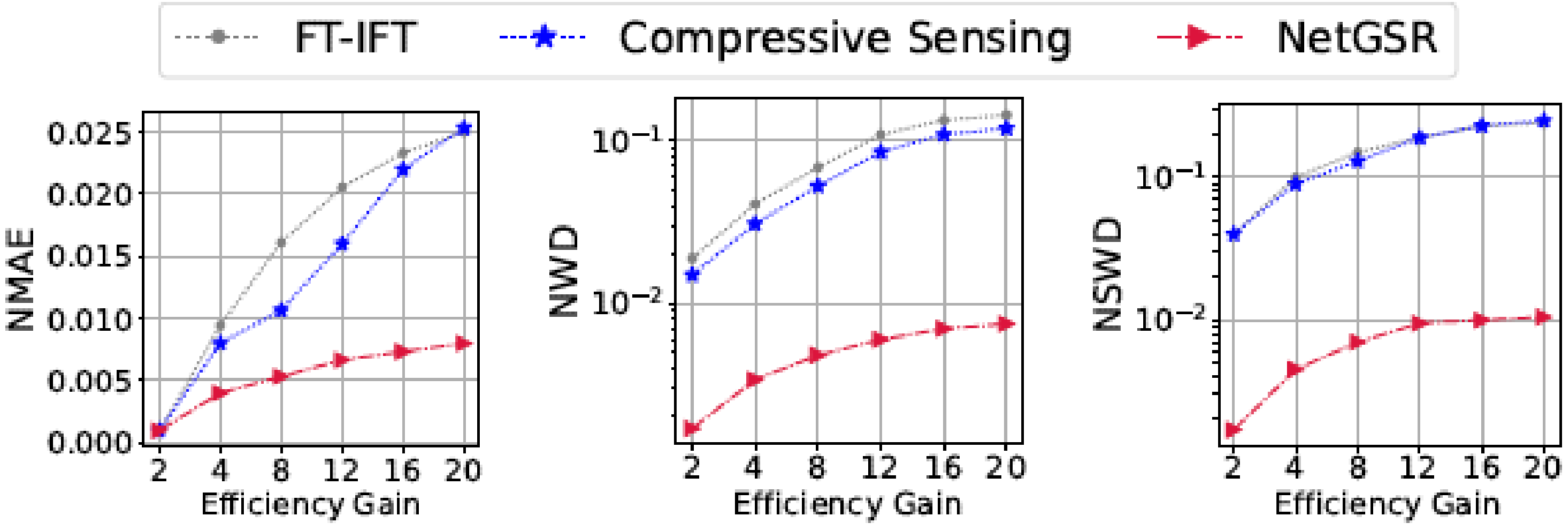
So, can we collect KPI data at desired time granularity without incurring the overhead?

Our Solution: NetGSR [CoNEXT'24]

- NetGSR = DistilGAN + Xaminer
  - DistilGAN – custom-designed generator to recover original high-resolution data stream given a low-resolution version
  - Xaminer for sampling rate adaptation



20x Efficiency Gain with 5G Open RAN KPIs



Summary

- Dramatic transformation of mobile networking system architecture and deployment in the past decade
  - Driven by operator economics
  - Key enabler: Disaggregation (at three levels)
  - Introduced new research challenges and opportunities
- Now amid another such transformation in the RAN towards data driven and AI powered operation
  - Similar motivation as before, more disaggregation and expand openness
  - New challenges to be tackled for this transformation to become a reality
- GenAI can be leveraged to address some of the key challenges:
  - Troubleshooting a disaggregated RAN - SpotLight [MobiCom'24]
  - Efficient acquisition of RAN data – NetGSR [CoNEXT'24]