O-RAN next Generation Research Group (nGRG)

Contributed Research Report

Digital Twin RAN: Key Enablers

Report ID: RR-2024-09

Contributors:

NVIDIA

Bell CA

CMCC

DeepSig

Dell

Ericsson

Nokia

Qualcomm

Rakuten

Verizon

VIAVI

Release date: 2024.10

Authors

Company	Name
NVIDIA	Lopamudra Kundu (**Editor-in-chief**) Xingqin Lin Emeka Obiodu
Bell Canada	Javan Erfanian
CMCC	Yuxuan Xie
DeepSig	Tim O'Shea
Dell	Hoda Dehghan Ibrahim Abualhaol Javad Mirzaei
Ericsson	Mirko D'Angelo
Nokia	Daiju Chiriyamkandath Antony
Qualcomm	Geetha Rajendran
Rakuten	Kexuan Sun
Verizon	Nirlay Kundu Vishwanath Ramamurthi Jin Yang
VIAVI	Chris Murphy Takai Eddine Kennouche

Reviewers

Company	Name
Reliance Jio	Vikas Dixit
Nokia	Niraj Nanavaty
DTAG	Jan Plachy
Capgemini	Avijit Manna
China Telecom	Zexu Li
China Telecom	Qingtian Wang
VTT	Tao Chen

Disclaimer

Copyright

Executive summary

Digital Twin Network (DTN) is foreseen as one of the essential tools for managing the complexity and demands for emerging 6G networks, offering a high-fidelity and real-time virtual environment that would mirror complex behaviors of the underlying physical network. This critical technology integrated with 6G networks will enhance network planning, monitoring, optimization, security, reliability and more. This, in turn, would enable telecommunication operators to improve the network performance, reduce the networks' capital and operational costs and accelerate innovation while maintaining high service quality and availability. Creating the perfect recipe of a highly accurate DTN requires several ingredients to come together into the mix - data collected from the physical network, advanced modeling techniques for DTN, and robust interfaces around DTN. The objective of this research report is to deep dive into a set of key technologies that would enable the realizations of these key ingredients into shaping up the high-fidelity digital avatar of a 6G network, and to ensure seamless interaction and synchronization between the digital and physical realms. In particular, this research report focuses on Digital Twin for Radio Access Network (RAN) or DT-RAN, one of the most complex domains within a network infrastructure. Understanding intricacies of Digital Twin enablers for RAN will be crucial in extending the scope of this technology beyond the RAN Domain and to other parts of the network.

In Section 1 of the report, we provide a general introduction of Digital Twin as a pioneering technology for 6G RAN, and a brief overview of the related works happening in industry standards fora. Section 1 sets the stage for the following Section 2 and the subsections therein, where we get into the detailed analysis of various building blocks of a DT-RAN and relevant technology enablers that are going to make these building blocks realizable in practice, along with a few illustrative use cases explaining how these technology enablers play a key role in developing DT-RANs catering to those scenarios. Section 2.1 focuses on explaining the importance of data for DT-RAN and various sources of data collection along with relevant techniques for data collection, generation, and augmentation. Section 2.2 addresses the modeling aspects of DT-RAN, highlighting the modeling of network elements, physical environment, network subscribers and ultimately, the end-to-end systems. Sections 2.3 and 2.4 highlight the importance of accelerated computing, virtualization techniques and trustworthiness management provisions to realize a highly efficient DT-RAN. The following Section 2.5 elaborates on aspects related to interfaces around Digital Twin for RAN, enabling efficient intercommunication and information exchange between various components of a DT-RAN as well as between DT-RAN and underlying physical networks. Section 3 concludes the research report with a summary of key findings from Section 2.

Table of Contents

List of abbreviations

3D	Three-Dimensional
3GPP	3rd Generation Partnership Project
ADASYN	Adaptive Synthetic Sampling
AI	Artificial Intelligence
API	Application Programming Interface
ASIC	Application-Specific Integrated Circuit
BLER	Block Error Rate
CAD	Computer Aided Design
CBRS	Citizen Broadband Radio Service
CCTV	Closed Circuit Television
CIR	Channel Impulse Response
CM	Configurations Management
CNF	Cloud-native Network Function
CNN	Convolutional Neural Network
cuBLAS	CUDA Basic Linear Algebra Subroutine
CUDA	Compute Unified Device Architecture
cuDNN	CUDA Deep Neural Network
cuFFT	CUDA Fast Fourier Transform
DLA	Deep Learning Accelerator
DPU	Data Processing Unit
DRL	Deep Reinforcement Leaning
DT	Digital Twin
DTN	Digital Twin Network
DT-RAN	Digital Twin for Radio Access Network
E2E	End-to-End
EM	Electromagnetic
EMI	Electromagnetic Interference
EVM	Error Vector Magnitude
FCAPS	Fault, Configuration, Accounting, Performance and Security
FH CUSM	Fronthaul Control-User-Synchronization-Management
FM	Fault Management
FPGA	Field Programmable Gate Array

FTRT	Faster-Than-Real-Time
gNB	Next Generation Node B
GAN	Generative Adversarial Network
GNN	Graph Neural Network
GPU	Graphics Processing Unit
HLS	High Level Synthesis
HO	Handover
HPC	High Performance Computing
IMU	Inertial Measurement Unit
IoT	Internet of Things
IRTF	Internet Research Task Force
ISAC	Integrated Sensing and Communication
ITU-R	International Telecommunication Union Radiocommunication Sector
ITU-T	International Telecommunication Union Telecommunication Standardization Sector
KPI	Key Performance Indicator
LiDAR	Light Detection and Ranging
LTE	Long Term Evolution
M2M	Machine-to-Machine
MAC	Medium Access Control
MAGMA	Matrix Algebra for GPU Multi-core Architecture
MDT	Minimization of Drive Tests
ML	Machine Learning
MLP	Multi-Layer Perceptron
NDT	Non-Deterministic Testing
NN	Neural Network
NPU	Neural Processing Unit
NR	New Radio
OAM	Orchestration and Management
O-CU-CP	O-RAN Central Unit – Control Plane
O-CU-UP	O-RAN Central Unit – User Plane
O-DU	O-RAN Distributed Unit
OpenACC	Open Accelerators

OpenCL	Open Computing Language
OpenMP	Open Multi-Processing
O-RAN	O-RAN Alliance
O-RU	O-RAN Radio Unit
OSS	Operational Support Systems
PDCP	Packet Data Ciphering Protocol
PDP	Power Delay Profile
PBCH	Physical Broadcast Channel
PDCCH	Physical Downlink Control Channel
PDSCH	Physical Downlink Shared Channel
PRACH	Physical Random Access Channel
PUCCH	Physical Uplink Control Channel
PUSCH	Physical Uplink Shared Channel
PII	Personally Identifiable Information
PM	Performance Management
QoS	Quality-of-Service
RADAR	Radio Detection and Ranging
RAN	Radio Access Network
RF	Radio Frequency
RIC	RAN Intelligent Controller
RIS	Reconfigurable Intelligent Surface
RLC	Radio Link Control
RNN	Recurrent Neural Network
RSRP	Reference Signal Received Power
RSRQ	Reference Signal Received Quality
RSSI	Reference Signal Strength Indicator
RT	Real-Time
SAS	Spectrum Access System
SDK	Software Development Kit
SINR	Signal-to-Interference-plus-Noise Ratio
SMO	Service Management and Orchestration
SMOTE	Synthetic Minority Over-sampling Technique
SoC	System-on-Chip

THz	Terahertz
TPU	Tensor Processing Unit
TTI	Transmission Time Interval
UE	User Equipment
USD	Universal Scene Descriptor
V2X	Vehicle-to-Everything
VAE	Variational Auto-encoder
VNF	Virtualized Network Function
Wi-Fi	Wireless-Fidelity

List of figures

1 Introduction

As the telecommunication experts across the academia and industry continue to shape the scope of the emerging sixth generation (6G) wireless networks, it has become apparent that the onset of 6G will manifest into many novel applications and services, ranging from multi-dimensional sensing, network and computing convergence, pervasive intelligence, immersive multimedia, extended reality, holographic communication to connectivity for industry 4.0 and beyond, as illustrated in *ITU-R M.2160-0*. Enabling this wide range of use cases requires addressing a diverse set of requirements, which would be difficult to meet with the previous generations of wireless networks. To that end, various state-of-the-art technologies have emerged as key enablers for 6G use cases, among which the Digital Twin (DT) has stood out as one of the highly promising candidates to facilitate the design, analysis, operation, automation, and intelligence of 6G wireless networks. It is noteworthy that the digital twin concept has been around for quite some time across other industry verticals. Internet Research Task Force (IRTF) is working actively on the exploration of digital twin as a rapid adoption technology for Industry 4.0 requirements. The technical draft [3] published by IRTF describes reference architecture and key challenges for building a network digital twin [3]. Similarly, digital twin architecture proposals for performance optimization of optical networks are published by IRTF in [4].

A Digital Twin Network (DTN) is a digital replica of a communications network, or part(s) of a communications network, including, for example, any combination(s) of physical network elements and components, virtualized/cloud-native (containerized) network functions (VNFs/CNFs), physical hosts for such VNFs/CNFs etc. Unlike conventional network simulators, the DTN supports communication between the physical network and the virtual twin network to achieve real-time interactive mapping. DTN provides an experimentation sandbox where network designers, developers, and operators can play with various network configurations and topologies in a risk-free environment, before applying those changes in the real network. Inspired by the great potential of DTs for wireless networks, several initiatives have emerged in standards bodies to develop initial guidelines for DTNs. The International Telecommunication Union Telecommunication Standardization Sector (ITU-T) has released a recommendation (*ITU-T Y.3090*) in February 2023 that describes the requirements and architecture of DTNs. The International Telecommunication Union Radiocommunication Sector (ITU-R) has published a report (*ITU-R M.2516-0*) in November 2022 on future technology trends of IMT systems towards 2030 and beyond with a list of key emerging use cases for 6G and listed DTNs as one of the important candidates in that category. In particular, ITU-R recommends a top-level design of the Digital Twin for Radio Access Network (DT-RAN) to be considered first, before extending the scope to beyond the Radio Access Network (RAN). Accommodating diverse physical RAN networks, DT-RAN can be designed as a first candidate for DTNs, as per ITU-R recommendation.

RAN is arguably the most complex and compute intensive part of a wireless network infrastructure. Creating a high-fidelity digital replica of RAN, which would truthfully mirror a slew of complex behaviors consistently and continually with time is a non-

trivial challenge. As RANs are evolving towards increased heterogeneity, dynamicity, and complexity, it is natural that DT-RANs would require advanced technical capabilities to faithfully replicate the physical network traits in the virtual domain. To that end, creating a high-fidelity and scalable digital replica of the real-life wireless propagation channel between the physical network and the UEs is also an important part of DT-RAN.

In the subsequent sections, we will deep dive into some of these key technologies that would be critical to enabling highly sophisticated DT-RAN. In particular, we will focus on O-RAN Alliance (O-RAN) specific aspects of these key enabling technologies that would make DT-RAN a reality for the open RAN ecosystem as it steps into the 6G era.

2 Key Enablers of DT-RAN

The foundation of a DT-RAN can be built upon the three key pillars of a general DT: Data, Models, and Interfaces. DT-RAN's ability to represent its physical counterparts with high fidelity hinges upon various characteristics of data used to construct its models, and the precision as well as accuracy of those models in emulating and predicting physical network's behavior. Meanwhile, interfaces surrounding DT-RAN facilitate its seamless interaction with the underlying physical network and related network applications, as well as between various modules of the DT-RAN itself. The inception of a DT-RAN begins with enabling various facets of these key pillars. Alongside, making DT-RAN efficient, reliable, and trackable are also crucial, which require high performance computing capability, ability to monitor DT-RAN's trustworthiness over time and visualize DT-RAN's complex actions and operations.

Therefore, creation of a high-fidelity DT-RAN requires a set of key enablers coming together, namely data, modeling, computing, visualization, trustworthiness management and interfaces around DT. The details of these various facets are illustrated in the following subsections, along with specific use cases illustrating how these key enablers can come together to realize DT-RAN for certain scenarios. In particular, the use cases addressed in the following subsections are derived from the research report published by O-RAN titled "Digital Twin RAN Use Cases" [42] and evolve around DT-RAN for AI/ML training, testing and performance assurance, network planning, network energy saving, site-specific network optimization and network automation.

2.1 Data: Acquisition & Management

Data is one of the fundamental building blocks of DT-RAN. In general, DT of a wireless network can acquire data from various sources as depicted in Figure 2-1. It is noteworthy that Figure 2-1 is showing an illustrative (and non-exhaustive) list of data sources for DT-RAN and there could be additional sources of data depending on use cases (for example, data from intelligent transportation systems, emergency, and disaster response systems etc.).The collected data is used for creating, calibrating, updating, and configuring a digital twin. Specifically for DT-RAN, its architecture and

interfaces are essential for enabling various aspects of data, including data acquisition and data management. O-RAN has established a cloud-native RAN architecture with open interfaces for generic, open RANs. Those can provide the basis for a viable solution for DT-RAN data collection and data management.

2.1.1 Data Acquisition: Collection, Generation & Augmentation

Data can be acquired by DT-RAN through various means, including data collection from real world (e.g., from the physical environment or from the network elements), synthetic data generation, and data augmentation - a hybrid approach of combining synthetic data with real data to create a richer dataset. The efficacy of a high-fidelity DT-RAN relies both on the quality as well as the quantity of data collected from these diverse sources and enabled by various means.

Acquisition of data from various sources as depicted in Figure 2-1 can be one-time, periodic, or event-based. For example, in case of a DT-RAN deployed in an urban city, Computer-Aided Design (CAD)/map data can be either a one-time input, or very low frequency periodic update (e.g., every 6 months), or an event-based update (when either an aerial imaging survey, or city planning indicates a change in the city landscape). On the other hand, camera, Light Detection and Ranging (LiDAR) or sensors can provide updates with a higher frequency (of the order of few minutes to few hours), capturing changes in the surrounding of these devices.

Data from Radio Frequency (RF) measurements, alarms and performance metrics can be used for calibrating the DT-RAN, or for assessing the impact of a ‘what-if’ study related to physical RAN configuration.

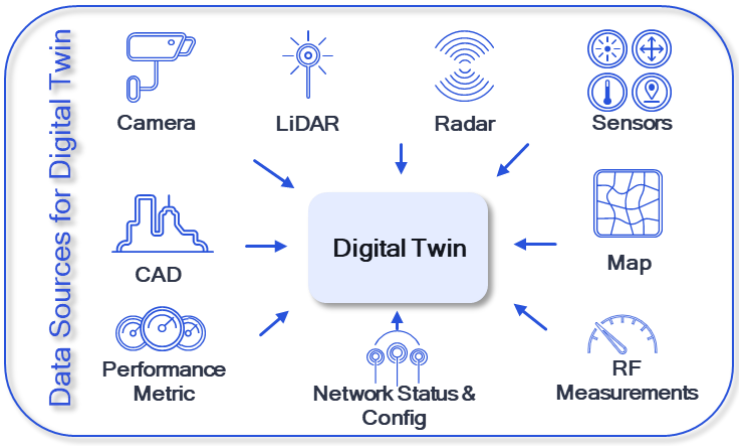


Figure 2-1: Example Data Sources for a Wireless Network Digital Twin

2.1.1.1 Collection of Data from Physical Environment

Data collection from the (complex) physical environment surrounding the live network is crucial for enabling high-fidelity DT-RAN that can replicate the commercial RAN deployment as closely as possible. Channel and propagation conditions can be

approximated by empirical models, but discrepancies with reality will mean that predictions of a digital twin will have errors. The error will lead to suboptimal network performance. Another confounding factor is that the physical environment is dynamic and constantly changing, with changes ranging from short term fluctuations to diurnal and seasonal cycles. Fading and shadowing are constantly fluctuating in unpredictable ways as objects and people move, leaf cover changes, weather changes, and structures are built, reconstructed, and demolished. These characteristics vary by frequency, adding further complexity to the physical environment modeling. The spatiotemporal and connection characteristics of the mobile network subscribers are also affecting the modeling of the network environment.

Despite the complex nature of the physical environment and its dynamically evolving characteristics, there exists various technology enablers that can facilitate data collection from various sources in the physical environment, with the required frequency and quantity, to create the necessary models for DT-RAN. Data collected from the physical environment capture various aspects of wireless systems. Type of parameters captured as part of the data collection depends on the layout/scenario of the physical network. For example, data collection for most of the public-outdoor networks may require collection of only coarse level parameters like Reference Signal Strength Indicator (RSSI)/ Reference Signal Received Power (RSRP)/Location Information or other parameters derivable from these (e.g., Reference Signal Received Quality (RSRQ)), whereas a private-indoor network may require collection of images/video streams from Closed Circuit Television (CCTV) cameras for three-dimensional (3D) model updates. In another example of some other remote outdoor condition, e.g., for over-the-sea maritime scenario, it may require to additionally collect more detailed information e.g., sea state information, weather condition etc.

Type of data that can be captured from physical environment and radio network can include but is not limited to:

- RF channel measurements (RSSI, RSRP, Signal-to-Interference-plus-Noise Ratio (SINR)).
    - More detailed information like Power Delay Profile (PDP)/Channel Impulse Response (CIR) for indoor scenarios.
- Throughput/Latency measurements.
    - Capturing user experience.
- Positioning information.
    - Geo-tagging radio data.
- Visuals (image, video, LiDAR scans) capturing objects/surrounding/scene.
- Integrated Sensing and Communication (ISAC), Radio Detection and Ranging (RADAR)/RF sensing captures (base stations/user devices acting as RADARs).
- Sea state information e.g., wave height for maritime scenarios.
- Weather condition e.g., rain rate, fog, snow, temperature, wind speed and atmospheric conditions.

The data can be used as ground truth for building and calibrating corresponding digital replica of the physical environment to be used as a part of DT-RAN. It can also be used for evaluating the performance of a DT-RAN, i.e., how accurate are DT-RAN predicted results.

Enabling Technologies:

Measurements of the physical environment surrounding a physical network can take place in a variety of ways. When the network is initially installed, the technicians who install and service the network will often collect measurements of the physical network. These measurements can include signal coverage and quality for different cells at different locations. While some of these measurements can become outdated and less relevant over time, they can be relevant in some circumstances. Measurement campaigns need not be associated with the installation and servicing of infrastructure. Although they can be costly, they can be performed at any time.

Sensors such as signal scanners or active mobile devices can be deployed in the network to collect ongoing measurements without the need for a technician to intervene. So-called smart sensors can make measurements much like those at the installation stage and may be installed on a fixed structure or attached to a vehicle to survey a wider area. The network itself can be used to make measurements of the physical environment. The base stations can collect RF measurements and mobile devices generally make measurements (e.g., RSRP/RSSI), reporting these to the infrastructure for radio resources and mobility management.

ISAC is gaining attention as a promising technology. The signals that are used for communication also have utility for sensing, where the nature of backscattered signal correlates with objects in the environment and doppler shift correlates with motion. This facilitates localizing objects in the environment or establishing information about their characteristics or identity, along with measuring information about environmental conditions. While this field is in its infancy, its potential for enabling digital twin as it matures is promising (for example, German national 6G Project on ISAC titled "*komsens*").

In general, data collection techniques from the physical environment can be broadly categorized into two buckets: legacy techniques and enhanced techniques. Following is a non-exhaustive list of key enablers corresponding to each of these technique-categories:

Legacy Techniques:

- Drive Tests with reference devices capturing RF/Positioning/Throughput/Latency measurements.
- Minimization of Drive Tests (MDT) based data collection in cellular networks from user devices capturing information supported as per standard procedures/protocols [1].

Enhanced Techniques:

- MDT Enhancements
    - Enhancing the legacy MDT to support additional requirements for digital twin, if needed.
- Smart Cameras (with On-device Artificial Intelligence/Machine Learning (AI/ML) support)
    - Integrating cameras with RAN deployment for capturing images/videos of installation sites. Collected data can be processed in pseudo real-time/non-real time to generate information on changes in the physical environment. The change information can be used to update the digital twin.
    - Example - in a public-outdoor network (like marketplace) camera data can help in detecting any changes required in 3D model, like new objects causing shadowing/blockage, or change in subscriber density (based of traffic flow). Similarly, in a public-indoor network (like stadium) camera data can help in updating 3D model with subscriber density. Updated digital twin can optimize network for the sporting events.
- LiDAR Scanners
    - 3D model creation/update of network deployment areas can be performed using data collected from LiDAR scanners (vehicle or drone mounted).
- Any other Internet of Things (IoT)/Sensor/ISAC devices.
    - IoT/Sensor devices (e.g., RSRP monitor, device density monitor, temperature/humidity sensors etc.) can be installed in the network deployment areas to do ‘autonomously monitored’ data collection and upload to digital twin. The data can be used for digital twin update/calibration.
- Maps
    - Digital twin can create/update 3D model of the network deployment areas based on the Topological/Topographical Maps.

To summarize, on top of the commercial RAN itself, it is essential to represent the physical world, topology, and objects accurately that could impact the radio access network. Thus, technologies like smart sensors, IoT devices, and ISAC will play an important role in collecting data from the physical environment and enable accurate modeling of these various entities in a radio access network deployment environment.

Example Use Cases:

Use Case 1: DT-RAN for Network Planning

Smarter planning decisions can be made if DT-RAN can accurately predict the propagation and channel characteristics between a transmitter and a receiver in a physical network. Given a proposed site for a new antenna, or a new frequency carrier at an existing site, a digital twin of the radio environment will be able to predict what signal strength and quality can be achieved for a given distribution of subscribers. Hence the data collection from physical elements can enable the use case of “DT-RAN for network planning.”

Type of data collection and its usage in enabling network planning include but are not limited to:

- Data collected from physical environment:
    - RSSI/RSRP measurements along with measurements providing Positioning information at selected locations in the physical area under planning using legacy methods of data collection as explained in the previous subsection.
        - Locations are selected carefully to ensure data is available for diverse set of channel conditions in the area under planning.
- Data generated from digital twin:
    - RSSI/RSRP heatmaps for all base stations within the area under planning.
    - Mapping of location information and heatmap, i.e., overlaying heatmap on the actual physical location vector.
- Calibration of digital twin using collected data.
    - Comparison between RSSI/RSRP values from digital twin generated heatmap and the values collected from physical environment.
        - Any differences between these values require calibration adjustments to be performed on digital twin. One or more aspects of the digital twin may require correction/update, for example -
            - *3D Model* – Refinement of surfaces, edges, objects, etc.
            - *Material Mapping* – Reassignment of materials to minimize the gap between electromagnetic and optical properties.
            - *Ray tracing & electromagnetic calculations* – Minimizing inconsistencies, configuration differences between physical and digital twin configurations.
- Iterating above steps with new measurement and location datasets to minimize the error between physical environment and digital twin predictions.

Once the DT-RAN has achieved required accuracy, it can be used to generate exhaustive RSSI/RSRP heatmaps for various candidate cell-site locations & beam configurations. Generated heatmaps can be used by AI/ML based use cases like coverage and capacity optimization, mobility optimization etc. to adjust and arrive at optimal network configuration, minimizing/eliminating coverage gaps.

Use Case 2: DT-RAN for Network Performance Predictions

The DT-RAN replicates the physical network's components and operational conditions, enabling an accurate analysis and prediction of the network performance before actual deployment. This predictive capability is invaluable for optimizing resource allocation, enhancing efficiency, and ensuring that the network meets performance expectations and service quality. It avoids trial-and-error approach in traditional network deployment, minimizes the risk of costly errors, and enhances overall network reliability.

Figure 2-2 shows how a DT-RAN can closely mimic a 5G RAN deployment based on O-RAN architecture and interfaces. The DT-RAN system here can include elements

in the O-RAN architecture like O-RAN Distributed Units (O-DUs), O-RAN Central Unit-Control Planes (O-CU-CPs), O-RAN Central Unit-User Planes (O-CU-UPs), O-RAN Radio Units (O-RUs) etc. Also, the DT representations of the network elements in the DT-RAN system can be implemented to support the same interfaces as the real network elements themselves. This includes 3rd Generation Partnership Project (3GPP) RAN interfaces like X2, Xn, NG, F1, E1 and O-RAN defined interfaces like the Open Fronthaul Control-User-Synchronization (FH CUS) plane, Open FH Management plane (M-plane), O1, O2, A1, R1, E2, Y1 etc.

In Figure 2-2, the F1 interface can provide information on radio environments, such as radio channel conditions and radio resource allocations. Thus, the DT-RAN can use data to emulate the radio access network performance and behaviors.

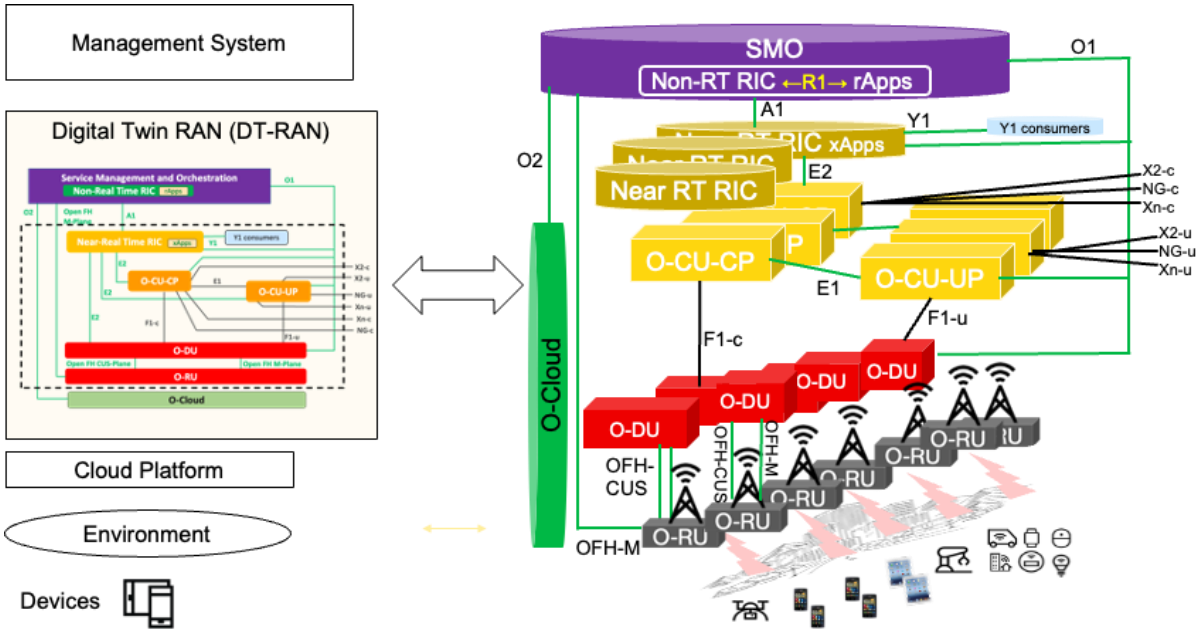


Figure 2-2: DT-RAN Architecture to Mimic Real O-RAN Deployment

In general, DT-RAN should be able to flexibly replicate different operator deployment scenarios including different topologies with their respective hierarchies of network elements. Novel modeling, data minimization and representation approaches can be used to minimize the amount of data transfer involved, striking a good balance between the data quality and the data quantity. The DT-RAN can be fully virtualized and run on a similar/same cloud platform that the operator uses for the rest of the virtualized O-RAN elements.

2.1.1.2 Collection of Data from Network Elements

In the previous subsections, network elements are described to have utility for measuring the physical environment, but their potential for enabling DT-RAN goes beyond that. A specific network element such as a physical or virtual network function can expose a certain functionality. This is typically in the form of the interfaces defined

by industry standards or specifications. Generally, a well-tested commercially deployed network function can be relied upon to conform to the standards. However, nodes generally contain proprietary algorithms, for example, algorithms for scheduling, prioritisation, encoding, dealing with congestion and impairments, etc. There can be wide ranges of responses that are conformant to the specifications. These responses in general vary by the situation in which the node finds itself. How it responds under congestion may be quite different to how it operates in lightly loaded conditions. Sparse measurements of a specific network element's behavior in response to extraneous stimuli can be used to build a digital twin, which can predict this element's behavior under a wide range of conditions. Examples of such extraneous stimuli may include impairments in network connectivity, network loading conditions etc.

In addition to extraneous conditions, a network element's behavior is also shaped by its internal state, which may include configured parameters, and the connection management process associated with that network element. In contrast to virtual network elements, physical network elements are affected by physical phenomena such as temperature. Virtual network elements can also have some performance coupling to the physical compute nodes on which they reside, and this can consequently manifest into the response of the virtual network element. Being able to model how the different components of the network, including the physical elements and virtual elements, the underlying physical compute nodes, and the transport connectivity between them will behave in given scenarios underpins the 'what-if' scenarios.

In general, DT-RAN is a collection of both physical environment as well as the radio access network side infrastructure of the mobile network operator. Data from both these entities is needed to create an end-to-end simulation of the real network. Data collection from the network elements is essential to provide insight into the network operation as explained above. In the context of an O-RAN architecture, data collected by performance and configuration monitoring systems through O1 interfaces can be utilized by non-Real Time (non-RT) RAN Intelligent Control (RIC) and near-Real Time (near-RT) RIC through interfaces designated in Figure 2-2. Those data can be used to optimize the network slicing and Quality-of-Service (QoS) of data flows.

DT-RAN has the potential to enable a service aware network for various vertical applications.

3GPP and O-RAN have defined various mechanisms for data collection from the network elements in a standardized way. The following are different standardised options currently available to collect data from RAN:

- Performance KPIs – defined in TS 28.522 and adopted in O-RAN O1 specification.
- Fault/Alarms – defined in TS 28.622 and adopted in O-RAN O1 specification.
- Trace mechanism – defined in TS 32.422 and adopted in O-RAN O1 specification.
- MDT Report – defined in TS 37.320 and adopted in O-RAN O1 specification.
- Configuration Parameters – defined in TS 28.531 and adopted in O-RAN O1 specification.

- O-RAN specified additional configurations, Key Performance Indicators (KPIs) and faults defined in O-RAN specifications.

These Fault, Configuration, Accounting, Performance and Security (FCAPS) data provide meaningful insight into network element configuration and its operation. Outside these parameters, the cloud related software configurations and deployment aspects are also specified in O-RAN O2 interface specification.

These data can be fed into DT-RANs. Configuration Management (CM) data contains information about the state of various parameters that are exposed by the network elements. These are generally available in the Service Management and Orchestration (SMO) or Operational Support Systems (OSS) databases, or via well-defined interfaces such as O2 as mentioned above. Fault Management (FM) data as mentioned above generally exposes occurrences of errors, impairments, anomalies and other unexpected or unwanted conditions. This can include overload conditions, situations where the network element has entered a recovery or failover state, failure of physical infrastructure or similar. This can reveal insights about the resilience or otherwise of the network element in the face of various problems. Performance Management (PM) data as mentioned above contains various KPIs, counters, statistics, etc. concerning how a node is used and the associated performance at various network layers. This can include throughputs, retransmissions count, statistics about modulation, coding, and channel state. It can also include numbers of attempts, successes, and failures for system accesses and connection events, along with numbers of normal releases and connection failures. Also, the proportion of physical resources that are used along with energy consumption can be captured.

In contrast to PM data, which is aggregated by network element, trace data as mentioned above is more granular down to the level of an individual User Equipment (UE) connection. This can include the sequence of messages and events that comprise of a specific traced call, along with information about Medium Access Control (MAC), Radio Link Control (RLC) and physical layer measurements. Network probes or data collection agents can collect data from the interfaces of physical or virtual network elements. These data may contain information to characterise the network element for creation of a digital twin such as the volumes of communication data to different endpoints or even finer granular data about the individual flows. Various RAN components can pass data through standardized interfaces shown in Figure 2-2. O-RU can pass data through front haul interfaces to O-DU, while O-DUs, O-CU-CPs and O-CU-UPs can pass data through O1 and E1 interfaces. SMO can use the data to oversee the orchestration, automation and control of RAN functions and infrastructure.

Enabling Technologies:

All the FCAPS data mentioned above are collected and processed using software tools like Prometheus, Grafana, Kibana etc. These tools can help collecting and processing the data and provide meaningful insights into the operation of the network.

Example Use Cases:

Use Case 1: DT-RAN for Testing of Non-RT RIC/rApps

The data collected from physical and virtual network elements can be used by developers/operators to test non-RT RIC applications (rApps) using the DT-RAN to increase the implementation maturity, performance, and confidence level of these applications before they are deployed in a real network. The DT-RAN should be able to interact with these applications just as it does with a real O-RAN deployment. For example, an rApp deployed in the non-RT RIC can interact with the DT-RAN the very same way it would interact with the real O-RAN deployment. Similarly, DT-RAN can be used for testing the efficacy of near-RT RIC applications (xApps) as well prior to deployment in real network.

Use Case 2: DT-RAN for Network Planning

The network element configuration parameters will provide the DT-RAN with the real time configuration of the network which will enable the DT-RAN to operate with the similar configuration. Once the replica of real network configuration data is fed to the DT-RAN, the configuration parameters can be iteratively tweaked and modified to achieve the required network planning. The performance and the fault data collected from the network can be used in the DT-RAN to understand the performance of the network with certain configuration. The physical environment data together with the data collected from the network elements can be iteratively used in the DT-RAN until the network planning objective is met.

Use Case 3: DT-RAN for Network Energy Saving

Data from network elements can contribute to network energy saving enabled by DT-RAN. For example, various data from the network elements including CM, PM and trace can be used to build models for how the energy consumption and performance are related to the configuration, loading and utilization of the radio access network. If this relationship is known, different configurations can be sought that deliver the services required while maximizing the magnitude of the energy saving.

2.1.1.3 Synthetic Data Generation

Data is required to train AI/ML models for specific use cases within the network or in the UE. These AI/ML models can be relevant both for the physical networks (e.g., AI/ML models deployed in non-RT RIC or near-RT RIC of O-RAN networks), as well as for modeling digital twins for various aspects/parts of the physical network (e.g. DT-RAN). AI/ML models trained for one specific use case, often cannot be used for another use case without additional training with additional data specific to the new use case. For example, a model trained for beamforming optimization for a cell-site in downtown San Francisco with data collected from that cell-site (or with data synthetically generated by simulating that specific cell-site in DT-RAN), will likely not be generic enough to be applicable for a cell-site in rural California. This increases the demand for site-specific data which is often not available in the desired volume needed to effectively train site-specific AI/ML models. One way to meet the data demand can be by using the DT-RAN of a specific cell-site to generate additional synthetic data. But to create a faithful digital replica of a specific cell-site itself, a large quantity of site-

specific data would be needed, which, in turn, would necessitate alternate ways for generating synthetic data for the DT-RAN first, before utilizing the DT-RAN for generating synthetic data for the AI/ML models deployed in the physical network.

The data necessary for training AI/ML models can also be of different types. For example, the data can include information about the demand placed on the network by the subscribers, the characteristics of the radio and other physical interfaces, and the configuration and state of the network elements. In some cases, the data may characterize the behavior of the network elements and their response to the stimulus. Sometimes, for purely software-defined network elements, the software itself may be used directly as a component of the digital twin. In other cases, where the element is partly comprised of hardware or where reproducing the complete functionality is not necessary and would increase the running cost of the digital twin, the behavior of the network elements may be modeled as part of the digital twin.

Historical data for all possible configurations, morphologies and scenarios would be best suited to train the AI/ML models for DT-RAN to understand the possible outcomes and model performance. However, it is not practical to have historical data for all permutations and combinations of scenarios and configurations. In the absence of sufficient historical data, data needs to be synthetically generated for specific scenarios and requirements to train the AI/ML models. Even when available, data is often costly to collect and manage, especially if it is dispersed throughout the network. Using limited data to train models can lead to models with biases and poor performance.

Synthetically generated data can augment authentic data and mitigate the impact of the lack of data, leading to better performing models with improved accuracy, precision, and recall. Synthetic data can express a wider range of scenarios than are experienced in a real network. For example, the demand placed on the network by the users along with the way that the various components of the network respond to this demand is highly multi-dimensional. Sampling the examples expressed in the real network will be only a subset of the possible range of combinations of stimulus and response. Synthetic data is a way to increase the sampling diversity across the whole space of scenarios. Then the models that are trained using this data will be exposed to a wider range of realistic scenarios.

Another motivation for synthetic data usage is to enhance privacy. Some data are granular and involve the behavior of actual subscribers as they use the network or the location in the network they are connected to. This can constitute personally identifiable information (PII) and so the data must be treated with care. This limits what data can be used for and what models can be trained with it. Synthetic data can address this issue by creating datasets that in aggregate have the macro characteristics of the users of the network, while in detail are entirely synthetic and not representative of individual users. This reduces restrictions on where the data can be used and what models can be trained with.

Finally, synthetic data generation can potentially be used to increase resilience to noise, missing features, and security threats. In general, wise use of data

augmentation can lead to increased model performance, generalization, and resilience.

Enabling Technologies:

DT in itself is a powerful technology for synthetic data generation. Data generation is one of the key use cases of DT-RAN. Once a DT of the real network is established, DT-RAN can be used to generate data to know the possible outcomes for different configurations and scenarios. DT-RAN can be used to generate synthetic data to train any AI/ML model for networks and UE by configuring the DT-RAN with the required scenario and generating the performance metrics.

In the context of generative AI, while transformers have become popularised by high profile models such as ChatGPT, they also have the potential to be used for synthetic data generation appropriate to mobile networks (e.g., synthetic channel data for target RF propagation scenario [43]), given appropriate precautionary measures are taken to combat detrimental issues such as “AI hallucinations”. As well as generation of natural languages, transformers can generate time series and event sequences, both of which can underpin aspects of a DT-RAN. A variational auto-encoder (VAE) is also capable of generating synthetic data. Once trained it maps the full-dimensional data into a probabilistic lower-dimensional latent space. Synthetic data examples can then be generated by sampling from the latent space and performing the corresponding decoding.

Generative Adversarial Network (GAN) is another example method to create a model capable of generating synthetic data. A *generator model* is created that can output examples in the form of the data for which synthetic examples are required. The generator in turn is trained by a discriminator. Various examples in the literature describe how GANs can be used to generate synthetic 5G data. The use of a GAN to generate synthetic examples of Call Detail Records (CDRs) is described in [45]. The performances of two types of GANs, viz. Conditional Tabular GAN and Topological Variational Autoencoder for generating synthetic 5G data are compared in [46].

Example Use Cases:

Use Case 1: DT-RAN for Network Planning

In sections 2.1.1.1 and 2.1.1.2, we have explained how data collected from real networks, including data from physical environment around the physical network and its network elements can be useful for expanding network planning for an existing site and predict network behavior and performance using DT-RAN for the existing physical network. However, physical data collected from existing network may not be adequate to accurately model and predict the behavior and performance of a network planned to be deployed in a different location. When location-specific data is lacking, synthetic data generated can be useful in filling the inadequacy of data availability, and help building a high-fidelity DT-RAN for the intended network site. The site-specific DT-RAN can be configured with the real network configuration and various deployment scenarios (e.g. different base station locations, antenna distributions, traffic models etc.), and the KPIs can be monitored to validate if the network planning objective is

met. Post network deployment, the configuration data generated out of the DT-RAN can be used in the real network and the KPIs from the real network can be fed into the DT-RAN in a loop achieving high accuracy and perfecting the analytics in DT-RAN.

Use Case 2: DT-RAN for Network Automation

The continuous testing envisaged by DT-RAN for network automation depends on the models that represent the full range of realistic scenarios. This is to ensure that the continuous testing based on the data from these models is fully comprehensive and results in robust network elements with maximal reliability. Such testing ultimately depends on presenting test vectors to the components or systems being tested, whereas these test vectors are typically highly complex. A basic approach is to collect data from the real network and replay it as test data. This can have value, but data taken from the actual network will not include scenarios that, while plausible or even likely, may not have occurred yet or are not captured in the test data collected from the network. In this scenario, synthetic data can be employed to create a set of test scenarios that is more representative of what will be encountered in the field.

This approach is not sufficient for the most reliable testing. It can be hard to classify the response to the test vectors as “right” or “wrong,” and thus “pass” or “fail” the test. It is the emergent behavior of the network and its associated performance metrics that ultimately matter. The grading of individual components as “passing” or “failing” a test becomes harder when the component is implemented with trained models in place of explicitly programmed algorithms. This is when the importance of digital twin becomes paramount.

The digital twin will be able to respond to the behavior of the system under test and characterise the resulting performance, thus grading the quality of the components making up that system. If that digital twin is powered by models that can create a wide range of synthetic data to challenge the system under test and respond to how the system under test reacts to those stimuli, the likelihood that any defects in that system or its components will be uncovered will increase. Consequently, the reliability of the system will be greater.

2.1.1.4 Data Augmentation

Data augmentation can be used to address some of the same challenges that can be addressed by synthetic data generation. These include increasing the volume of limited data and the need to respect privacy when data potentially contains PII. There are complimentary ways that data augmentation can be used alongside synthetic data generation, in the context of digital twin.

As collection of telemetry is not critical to delivering a service, it is often a lower priority and can sometimes get lost due to overload or congestion. Network operators can resort to data augmentation to impute missing data and mitigate the impact of data loss in the telemetry statistics. Sometimes models can be biased because certain regions of continuous feature or target space are more common than others, or certain

classes of discrete data dominate other classes. Data augmentation can offer some potential solutions to this imbalance problem. Model overfitting is a perennial problem for training models which can sometimes be mitigated by data augmentation. This helps in increasing the variation in the training data.

Data augmentation can help in other ways to build better models. In general, it is desirable to have models that embody feature invariance, i.e., the models are robust to transformations. For example, it is expected that a cell in one location experiencing a given set of stimuli and configuration to behave in a way similar to another cell experiencing the same set of stimuli and configuration but located in a different place. Data augmentation can potentially help to achieve feature invariance.

Finally, data augmentation can potentially be used to increase resilience to noise, missing features, and security threats. In general, wise use of data augmentation can lead to increased model performance, generalization, and resilience.

Enabling Technologies:

There are various data augmentation techniques useful for DT-RAN. Some of these techniques can benefit from domain-specific transformations. As one example, in the case of processing image data, data augmentation can be achieved by performing geometric transformations (e.g., rotation, zooming, cropping). As another example, in the case of timeseries and signal data, data augmentation can involve various transformations such as warping, windowing, jittering, and shifting.

Data imputation using techniques as simple as interpolation can also be valuable to augment data. Suppose the network architect needs a dataset of performance KPIs taken every minute, but the underlying system generating the KPIs systematically fails to deliver them at all timesteps. This will lead to a dataset with irregular timesteps. Interpolation can be employed to fill in the missing timesteps in this case, generating a more structured dataset that is useful for further machine learning model training.

Data perturbation and noise injection are also commonly used data augmentation techniques that are particularly useful for building resilient machine learning models that generalize well and can be invariant to data perturbations found at the time of deployment. Additionally, sample imbalance can be addressed with resampling augmentation techniques such as Synthetic Minority Over-sampling Technique (SMOTE) and Adaptive Synthetic Sampling (ADASYN). These various data augmentation techniques lead to the generation of new samples from a given original dataset, resulting in the creation of a larger “augmented” training dataset.

Other data augmentation techniques can enable the generation of additional features for datasets without necessarily increasing the sample size. For instance, extracting frequency domain representation from spectrum data and appending it to the time domain features. Another example could be extracting relative time from the absolute timestamps, and velocity from user mobility data to enrich a dataset containing only timestamps and coordinates. Advanced machine learning techniques can also enable data augmentation. Generative models, such as VAEs, GANs, and Diffusion Models

can enrich an existing dataset by generating new synthetic samples that are similar in whatever aspect deemed necessary for the downstream tasks.

Example Use Case:

Use Case 1: DT-RAN for AI/ML Training, Evaluation and Performance Assurance

Data augmentation can be important to enable DT-RAN for AI/ML training, evaluation, and performance assurance. This use case underscores the challenge of data availability and the need to assure the performance of the models trained on these data. Careful data augmentation can clearly work to mitigate problems with the lack of data availability. But it can also help to address the model performance assurance issue, as augmentation can ensure that more quantities and variety of test data are available. This can help to unearth excessive biases or variances in those models, and trigger model retraining as needed, resulting in higher model performance.

2.1.2 Data Management

As explained above, availability of rich and diverse dataset is the cornerstone of a high-fidelity DT-RAN. Once data is collected, its lifecycle management including data analysis, storage, retrieval, maintenance and termination (including data deletion and discarding) becomes an important aspect of overall DT-RAN's function. The efficacy of the underlying data management framework determines DT-RAN's long-term stability, performance, adaptability, security, and reliability. One of the key challenges that data management system of a DT-RAN can face is data drift, which is discussed in detail in the following subsection.

2.1.2.1 Data Drift Management

DT-RAN should behave as close as possible to the real system that it represents, e.g., a single network function within RAN, set of network functions within RAN, or entire RAN, etc. However, it is usually not straightforward to create an identical replica of the real system's object with all their dynamicity.

The DT-RAN can be realized using different types of network data, e.g., historical data, real-time data, near-real time data or any combination of such data, as well as using different data modeling framework (classical simulators, machine learning algorithms, statistics, etc.). Thus, the difference between the DT-RAN and the real-world object that it represents can largely vary based on the data used to realize the DT-RAN and the frequency with which the DT-RAN is being updated with the data from the real system. Correspondingly, the performance of the DT-RAN depends not only on the properties of the decision logic applied within DT-RAN, but also on the characteristics of the data used to realize the DT-RAN as well as the measure on how close the DT-RAN data represents the real system. Such measure can be expressed in terms of the data drift, i.e., the difference between the distribution of the real system data and the data of the DT-RAN.

Enabling Technologies:

In order to measure how close the DT-RAN represents the real system; different metrics need to be defined and collected. Such metrics can refer to the DT-RAN input or output data drift. DT-RAN input data drift is the drift in data used to realize the DT-RAN compared to the real system. It can occur due to the change in the data distribution of the real system, compared to the distribution of the data used for network modeling within DT-RAN. In such a case, the model of the network in DT-RAN will deviate compared to the actual network. Consequently, the output of the DT-RAN, e.g., predictions on the network state can deviate compared to the actual network state, i.e., there can be an output data drift. This can, in turn, impact the performance of the DT-RAN (e.g., accuracy or reliability of the recommendations and actions derived by the DT-RAN).

Therefore, the information on DT-RAN data drift, i.e., the changes in the data distribution (either as input or output of the DT-RAN) compared to the real system data need to be collected and monitored. Based on the collected data drift information, relevant actions can be performed, e.g., updates of the DT-RAN with the new data collected from the real system.

Example Use Case:

Use Case 1: DT-RAN for Network Predictive Analysis and Optimization

One of the common intended usages of DT-RAN is the so called ‘what if’ types of analysis and network optimization, where certain conditions and actions are evaluated in DT-RAN environment before applying the changes to the real system. In a specific use case, e.g., *mobility management***,** such ‘what-if’ evaluation would be useful for optimization of Handover (HO) parameters settings. In such use case, the DT-RAN would provide a digital replica of gNB configurations and environment model, where, for example, the ML approach is applied to decide on the best HO decisions. The ML model would be running in such a DT-RAN setup and would provide inference results with some precision/confidence. The inference results may be, as an example, the recommendation on the best target cell for performing HO to optimize mobility management. This can be evaluated in DT-RAN and in case the evaluation results in a good network performance, the corresponding configuration (e.g., HO settings to perform HO to recommended target cell) may be applied in the real system. However, when applying the HO settings to actual/real systems, the consumer of DT-RAN output needs to consider not only the ML KPIs (precision/confidence) but also the performance of the DT-RAN in which the ‘what-if’ analysis has been conducted. In other words, it is needed to have the information on how “close” to the real system the DT-RAN is in practice, to be able to apply with a certain confidence level the results obtained in such DT-RAN. In this concrete example, the ML model in the DT-RAN may use different input parameters such as performance measurements and UE measurements (MDT data) to predict the UE location and recommend the best target cell. The DT-RAN input data drift would refer to the difference in the data distribution between the parameters used by the ML model and actual MDT data collected in the real system. The DT-RAN output data drift may refer to the difference in the data

distribution between the predicted UE location and the actual UE location obtained from MDT measurements of the real system.

In the case when the DT-RAN data drift is low (or below required level), it implies that the recommendations on optimal HO parameters provided by the DT-RAN may be used in the real system with high confidence.

2.2 Modeling

The second pillar of DT-RAN constitutes “Models”. Building a comprehensive digital twin model for capturing the physical reality of a wireless network is of paramount importance for building DT-RAN. This requires developing faithful 3D models of the surrounding physical environment as well as the models for representing the wireless system existing therein. 3D models capture the nuances of the physical world where the wireless systems are operated and used for building the radio channel models. In general, wireless system is represented with a set of models corresponding to, for example, user device, network elements, traffic patterns [2] and applications etc., as shown in Figure 2-3. These models can be built independently and later can be integrated in a plug-and-play way as per the objective of the study/simulation.

Figure 2-3: Generic Illustration of Digital Twin Models for Wireless

Modeling of DT-RAN is multi-dimensional – comprising of various aspects of the physical counterpart that the DT-RAN needs to faithfully represent. It involves creating one or more dynamic and detailed virtual models that accurately reflects the status, configuration, statistical distributions, and performance of the physical RAN components, often in real-time. This virtual representation is continuously updated with data collected from sensors, monitoring devices, and other sources within the actual network. Some of these modeling aspects are addressed in the following subsections.

2.2.1 Modeling of Network Elements

Digital representation of a wireless network can consist of twins of one or more network elements depending on the scope/purpose of the analysis. These network element digital twins can represent complete functionality or part of the functionalities supported by real network element. For example, a DT-RAN used for the purpose of coverage/blockage analysis may only require gNB/UE digital element twins to support functionalities related to the broadcast of common/pilot channels on gNB and common channel reception/measurement on UE. Alternatively, a DT-RAN used for the purpose of capacity prediction would require gNB/UE digital element twins to support

functionalities related to CSI measurement/reporting, rate adaptation, power control, etc.

Network element digital twins can be realized either as simulator, or emulator, or AI/ML based models or any combinations thereof depending on the scale and the complexity of the functionalities represented by the element twin. Alternatively, multiple network elements/protocol layers can be represented by a single twin. These digital twins should support interfaces for configurations, control, and services like the real network.

Enabling Technologies:

One of the promising techniques for modeling network functionalities is Graph Neural Networks, detailed in Section 2.2.4. One can build scalable AI/ML models for network functionalities using these techniques. These models can be trained using synthetic data and tuned using field data, reducing the need of large volume of field data collection.

2.2.2 Modeling of Wireless Propagation Environment

As a key objective of the DT-RAN is to create a virtual replica of a real-world RAN to faithfully simulate wireless signal propagation, high-fidelity wireless signal propagation models are fundamental building blocks for the modeling of DT-RAN [5]. The wireless signal propagation modeling helps simulate how radio signals travel through the environment, interact with obstacles, and affect the coverage and performance of the RAN. Key aspects of the wireless signal propagation modeling include path loss modeling, shadowing and fading, obstacle penetration, frequency-dependent effects, antenna characteristics, terrain and environmental conditions, etc.

Wireless signal propagation modeling is crucial for DT-RAN in predicting the coverage area of radio signals within the network. By understanding how signals propagate in different environments and through obstacles, network operators can optimize the placement of antennas and plan for sufficient coverage using DT-RAN. Specifically, the DT-RAN with high-fidelity wireless signal propagation modeling allows network engineers to simulate different scenarios, such as the deployment of new base stations or the introduction of new frequency bands, before making changes to the actual network. This aids in efficient resource allocation and network expansion. Also, the DT-RAN with high-fidelity wireless signal propagation modeling helps in analyzing and mitigating interference issues within the network. By simulating the propagation of signals from neighboring cells or adjacent frequency bands, network operators can utilize the DT-RAN to identify potential interference sources and take preventive measures to ensure optimal network performance.

Enabling Technologies:

The physics of electromagnetic (EM) wave propagation is characterized by Maxwell’s equations. Solving Maxwell’s equations requires accurate knowledge of the boundary conditions at the interfaces between different materials, as well as the EM properties of these materials, which can be challenging to characterize accurately. To address specific aspects of wireless propagation without solving the full set of Maxwell's

equations, stochastic channel models and ray tracing are both techniques used in the field of wireless communication to model and simulate the behavior of radio waves.

Motivation for Deterministic Radio Propagation Modeling:

Stochastic channel models take a statistical and probabilistic approach [6]. They use statistical processes to model the random variations and uncertainties in the wireless channel. While stochastic channel models may capture statistical properties, they do not represent the precise geometric details of the environment or specific signal paths, leading to a lack of physical consistency and site-specificity. The shortcomings of the 3GPP stochastic channel models are discussed in [7]-[12] and summarized below. Take Integrated Sensing and Communication or ISAC as an example. The 3GPP stochastic channel models do not address target modeling and sensing, background environment modeling and differentiation from targets. To support ISAC study, it is critical to model sensing targets and background environment, including RADAR cross-section, mobility, clutter, and scattering patterns. Furthermore, the modeling must be spatially consistent, which is lacking in the stochastic channel models. Take Reconfigurable Intelligent Surface (RIS) as another example. Since RIS alters its reflected radio signal characteristics, the effects need to be captured by the modeling of wireless propagation. But the 3GPP stochastic channel models do not support the modeling of such effects. As another example, larger antenna arrays are expected to be used in the new spectrum such as 7-24 GHz and sub-THz bands, calling for additional considerations such as near-field effects of the channel, spatial consistency between a device and different radio units, different channel blockers (e.g., doors, wall, windows, foliage, concrete, and cars), spherical wave propagation, and spatial non-stationary effects of the channel. However, stochastic channel models often assume spatial stationarity. Furthermore, the use of new 7-24 GHz spectrum requires channel modeling consistency across different frequencies, which is yet to be validated in the current 3GPP stochastic channel models. Last but not the least, AI/ML is becoming an essential feature of 5G-Advanced toward 6G. Using data generated by stochastic channel models to train and test AI/ML models will lead to overly optimistic assessment of the models and result in failure of the AI/ML models when deployed in real networks.

In short, wireless networks are becoming increasingly heterogeneous, encompassing different inter-site distances, antenna array dimensions and makeup, radiated powers, frequency bands, to name a few. Correspondingly, the wireless channel modeling needs to provide consistency and, above all, a correct representation of the frequency, spatial, and temporal correlation across base stations and devices. Achieving this without a propagation model grounded on the underlying physics of the scattering phenomena is simply unnatural, prone to modeling error and possibly a huge waste of effort for the industry. These considerations call for deterministic, physics-based modeling for wireless propagation, especially ray tracing.

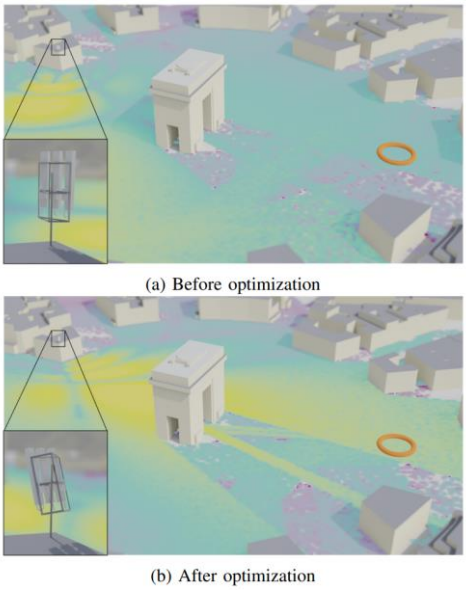
Ray Tracing Based Channel Models:

Ray tracing is a rendering and simulation technique used in computer graphics, optics, and other fields to simulate the way rays of light or other radiation travel through a

virtual environment. In the context of wireless communication and radio wave propagation, ray tracing is often employed to model and simulate the paths that EM waves take as they propagate through various materials and interact with surfaces and obstacles [13]. It provides a deterministic and physics-based modeling approach that simulates the paths of individual rays of EM waves, considering reflections, refractions, diffractions, and other interactions with objects and surfaces. It offers high-resolution simulations, capturing the specific paths of rays and the effects of the surrounding environment, making it valuable for site-specific planning, antenna design, and network optimization in the field of wireless communication. Wireless ray tracing, when integrated into a DT-RAN, enhances the accuracy and effectiveness of network planning, design, and optimization. It enables operators to make informed decisions based on detailed insights into the radio wave propagation environment, contributing to the overall performance and efficiency of the RAN.

Differentiable Ray Tracing Based Channel Models:

Differential ray tracing brings a unique and powerful capability to wireless signal propagation modeling for a DT-RAN. This technique combines traditional ray tracing methods with differentiability, allowing for the optimization of model parameters using gradient-based optimization algorithms, such as backpropagation. The concept was originally proposed for image rendering by computing the derivatives of a rendered image with respect to the scene parameters (e.g., scene geometry and materials) [44]. In the context of DT-RAN, differential ray tracing can be used to compute the derivatives of the functions of the simulated DT-RAN, such as coverage maps, with respect to the key parameters that impact the functions, such as material properties, array patterns, and geometries. Figure 2-4 shows an example of gradient-based optimization of the orientation of a transmitter using differential ray tracing in Sionna RT [14].



(a) Before optimization

(b) After optimization

Figure 2-4: Gradient-Based Optimization using Differential Ray Tracing [14]

In addition, differentiable ray tracing can be integrated into neural network architectures. By making the ray tracing process differentiable, it becomes possible to train neural networks to learn the complex relationships between input features (such as environmental conditions) and output signals (such as received signal strength or quality). The integration of ray tracing into AI/ML frameworks facilitates the joint training of AI/ML models and the optimization of signal propagation models, leading to more accurate and adaptive DT-RAN.

As another example, differentiable ray tracing may be adapted to model the effects of weather conditions on radio wave propagation. The modeling may incorporate weather conditions (e.g., rain, fog, snow, and atmospheric conditions) into the ray tracing framework. Making the ray tracing process differentiable with respect to weather parameters may allow for the computation of gradients that show how the weather parameters affect signal strength and quality.

3D Maps Based Channel Model:

3D map serves as a foundational element for wireless signal propagation modeling in a DT-RAN. It provides the necessary spatial context for accurate simulations, allowing the DT-RAN to capture the complex interactions between radio waves and the physical environment. A 3D map includes detailed terrain information, helping to model the topography of the landscape accurately. Terrain features, such as hills, valleys, sea state and uneven surfaces, significantly influence signal propagation. The 3D map enables the DT-RAN to account for these features, improving the accuracy of coverage predictions. 3D maps also include information about the height, location, and materials of buildings within the network area. This data is crucial for simulating the interaction of radio waves with buildings, considering reflections, diffraction, and shadowing effects. In conjunction with ray tracing techniques, a 3D map enables detailed simulations of how radio waves interact with the environment. In particular, the material properties influence how radio waves interact with various materials and structures in the environment. Understanding and accurately modeling material properties is essential for predicting signal behavior in the DT-RAN.

3D maps of real-world layout/scenario can be created using digital 3D models, generated by state-of-the-art computer vision and computer graphics technology. Typical source of inputs for building 3D models are LiDAR scans, Camera Images/Videos (2D, 3D, 360°), Inertial Measurement Unit (IMU) data, Location data, CAD models, Floor Plans, Topological/Topographical Maps and Satellite/Aerial Imagery. Generated 3D models need to be processed further for surface/object identification and material tagging; ensuring assigned materials are same/similar in electromagnetic properties to the corresponding real-world objects.

Computational complexity of running the ray tracing algorithm on such a model depends on the number of polygons (faces) in the model. Larger the number of polygons, higher is the compute resource and time requirements. Classical ray tracing techniques involve simulating real world phenomenon of reflection, refraction, diffraction, scattering and pass through. A ray path traversed between a transmitter

and a receiver can experience one or more of these phenomena. Phenomena of diffraction and scattering create many new ray paths from a single incoming ray.

To limit the size of a 3D model to a reasonable number of polygons (faces), one must do model simplification considering aspects like ‘size of the real-world layout/scenario’, ‘required fidelity of the model’, ‘required accuracy of the results’, ‘available compute resources’ and ‘type of interaction between real network and digital twin network (real-time, pseudo real-time, offline)’. For example, a 3D model of a small-scale indoor factory/warehouse may need to have large number of polygons (high fidelity), if the results are expected to capture exhaustive list of multipaths from ray tracing. In contrast to that the 3D model of a city urban area (large-scale outdoor) may require comparatively small number of polygons, if the results are expected to capture coarse level multipaths, blockages and shadowing.

Simulating radio wave propagation (RF channel) in the generated 3D model requires evaluating electromagnetic properties over the multipaths generated from ray tracing [15]. Alternatively, one can build and use AI/ML based models for predicting the RF channel behavior. Building of AI/ML models for predicting RF channel behavior based on object, geometry, and material is an area of interest for further research.

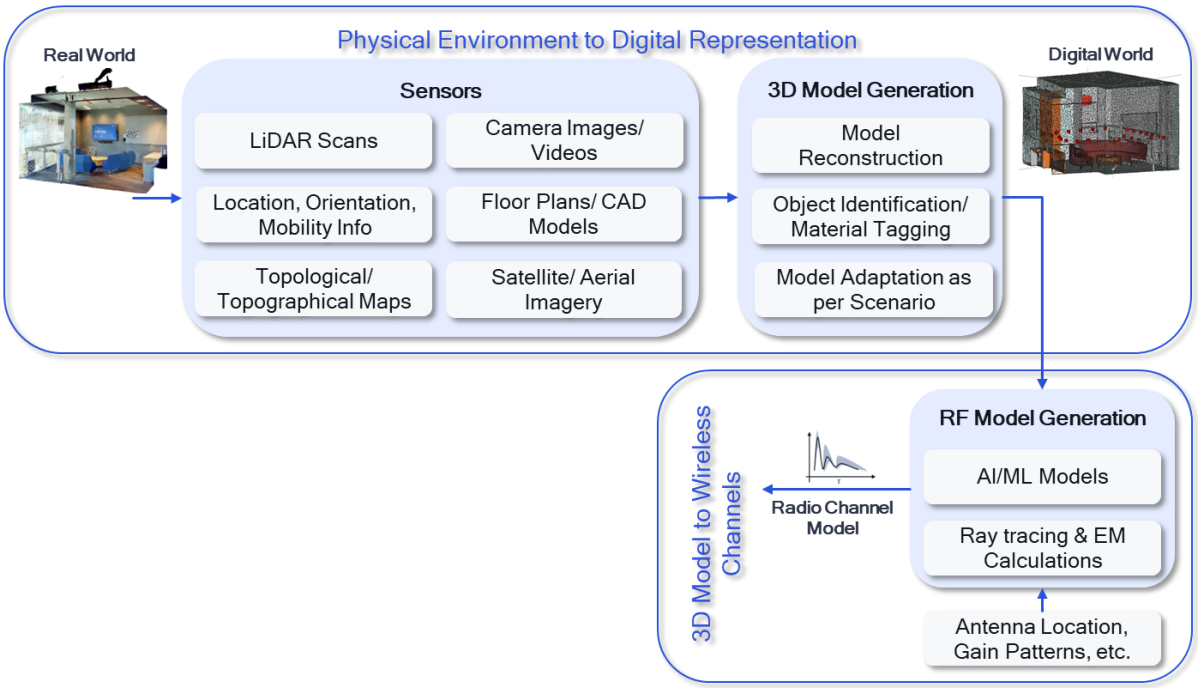


Figure 2-5: Modeling Propagation Environment

As one illustrative example, 3D models for “Conference Room Layout” are shown in Figure 2-5. Figure 2-5 was created using commercial LiDAR & Camera sensors, and 3D model reconstruction software. Radio channel simulation was done using wireless channel modeling software.

Generative AI Models:

Generative AI (GenAI) models can also be used to model statistical aspects of the wireless channel in a relatively lightweight way in addition to explicitly defined statistical channel models and full-fledged ray-tracing models. GenAI models are often an alternative (or complementary) to the stochastic channel models. While they are stochastic channel models in nature, they are typically not the explicitly defined stochastic channel models (e.g. TR 38.901 channels, etc.) but are learned neural networks which emulate the channel behavior from various data distributions. These models can help create a DT-RAN representing key aspects impacting wireless performance. GenAI models including GANs, Autoencoders, VAEs, and other such neural networks can be used to model aspects of signal propagation such as the power delay profile and multi-path fading statistics for an environment, modeling and reproducing the related probability distributions of a local environment with high accuracy. Figure 2-6 illustrates distribution matching (i.e. calibrating these GenAI models) from a single sector of over-the-air measurement data characterizing a channel impulse response in a macro-cell in Arlington, VA as compared to a generative autoencoder model which has been trained to accurately emulate statistical distributions of the aggregated channel response for one sector. Such a model can be trained with a relatively small volume of channel state data measurements, can generalize very well to the real underlying distribution, and can be used for a wide variety of DT-RAN applications which might not require a fully detailed spatial model of an area. In other examples, similar sorts of parameter learning can also be used in conjunction with ray tracing, for instance to help represent numerous local objects, properties, and effects.

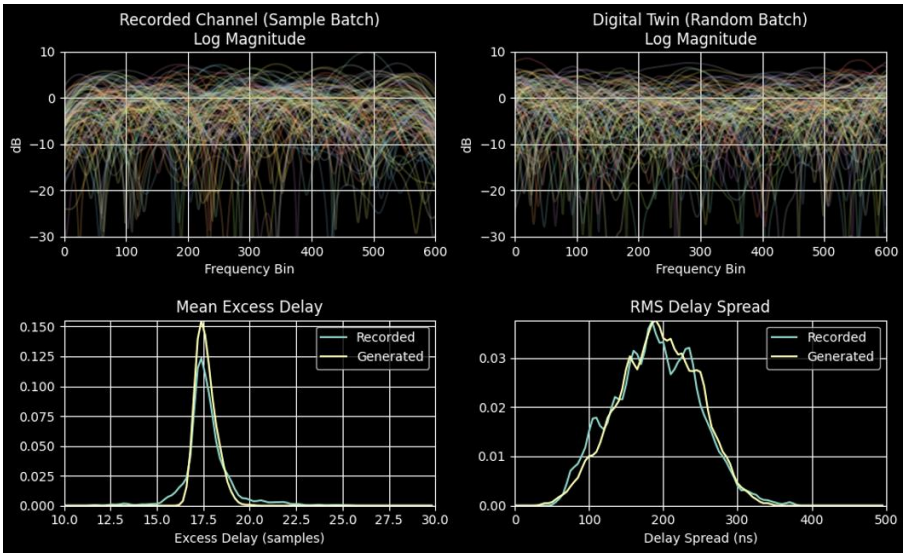


Figure 2-6: Generative-AI Based Impulse Response Modeling using Digital Twin

Example Use Cases:

Use Case 1: DT-RAN for Site-specific Network Planning and Optimization

Site-specific network planning involves leveraging detailed simulations of radio wave propagation for deploying network infrastructure components, such as base stations

and antennas, to achieve optimal coverage, capacity, and QoS in a specific geographical area.

To this end, the first step is to create a high-fidelity representation of the environment in the DT-RAN, including detailed 3D models of buildings, streets, and any structures that can impact signal propagation. Then wireless ray tracing can be implemented to model the propagation of radio waves in the target environment. Afterwards the interaction of EM waves with buildings, terrain, vegetation, and other objects can be simulated to capture realistic signal propagation characteristics. The accuracy of the site-specific network planning simulations needs to be validated by comparing the results with real-world measurements. The calibration ensures that the simulated environment in the DT-RAN closely mirrors the actual propagation conditions. In addition, visualization tools within the DT-RAN to present site-specific planning insights can be developed. The visualization can be used to show signal coverage maps, interference patterns, and other relevant information for better decision-making.

The ray tracing-based DT-RAN can be used for various site-specific network planning as well as optimization purposes, such as antenna placement optimization, frequency planning, capacity planning, and interference analysis. Take antenna placement optimization as an example. The ray tracing-based DT-RAN can be used to analyze how different locations and orientations of antennas affect signal coverage and quality, and evaluate the impact of obstacles, reflections, and diffractions on the propagation paths to and from the antennas. Such analysis and evaluation can be utilized to determine the optimal tilt angles, azimuths, and beamforming configurations for antennas to maximize coverage and signal quality and minimize intra/inter-cell interference.

In addition, data-driven GenAI models can be leveraged to model wireless propagation statistics such as delay spread, doppler, angular, or interference distributions in aggregate from cellular measurement data, and utilized for site-specific network optimization. AI/ML models within DT-RAN to simulate the PHY, MAC, and RRC, for example, can be simulated and optimized to tune for these channel statistics and spatial distributions within a sector to create locally optimized models which improve performance. In some applications processing stages, transmitter and receiver functions in portions such as Physical Random Access Channel (PRACH), Physical Uplink Control Channel (PUCCH), Physical Uplink Shared Channel (PUSCH), Physical Downlink Shared Channel (PDSCH), Physical Downlink Control Channel (PDCCH), Physical Broadcast Channel (PBCH), or others can be optimized based on KPIs such as Block Error Rate (BLER), Error Vector Magnitude (EVM), SINR, etc. The optimization is performed in the context of these wireless channel distributions to maximize processing performance. In some cases, this can further increase capacity, range, resilience, and efficiency of the RAN. By leveraging GenAI models to model channel data distributions, small and/or sparse data sets can often be extrapolated into smooth and complete distributions, stored compactly, shared between network elements such as RIC applications efficiently, rapidly used for training and simulation, and used across many network sectors and location to perform model management and optimization. In this case, RIC is envisioned as one platform which might host the

DT-RAN as an xApp or rApp, and provide various model optimization services such as the tuning of neural receivers or other signal processing functions.

In summary, site-specific network planning and optimization, empowered by wireless ray tracing and GenAI models in a DT-RAN, would allow network operators and planners to make informed decisions about the deployment and configuration of the network infrastructure. It has the potential to contribute to the efficient use of resources, improved coverage, and enhanced overall network performance in specific geographical areas.

Use Case 2: DT-RAN for AI/ML Training

A DT-RAN with a high-fidelity wireless propagation environment model can be effectively utilized for AI/ML training, allowing it to learn and adapt to complex radio wave propagation scenarios.

One can utilize the high-fidelity environment model to generate a diverse dataset that simulates a wide range of radio wave propagation scenarios within the targeted environment. This dataset can include variations in terrain, building structures, material properties, and other relevant factors. The DT-RAN can further assist in labeling the generated dataset with ground truth information, including expected signal strength, interference levels, and other relevant metrics. The DT-RAN with a high-fidelity wireless propagation environment model provides accurate reference data for training the AI/ML models (such as AI/ML based CSI feedback, beam management, and positioning), including the ground truth against which predictions are evaluated.

One may further integrate differentiable ray tracing techniques into the AI/ML models. This enables the models to learn from the simulations performed by the high-fidelity model and adapt their parameters based on the differentiability property. The integration supports end-to-end learning and optimization, where the AI/ML models learn directly from the high-fidelity simulations. This allows the models to capture complex relationships between input features and desired outcomes, facilitating more accurate predictions and optimizations within the DT-RAN.

2.2.3 Modeling of Network Subscribers

A typical network subscriber can be represented as a combination of a User Equipment, Subscription, Services and Application Data. Subscribers interact with network for exchanging control/user plane signaling/data.

Network subscriber(s) for a DT-RAN can be modeled using one or more inputs including but not limited to:

- Device Capability Model
    - Model capturing the capability (as defined in standards) of subscriber devices in terms of communication technology (Long Term Evolution (LTE)/ New Radio (NR)/Wireless-Fidelity (Wi-Fi)/Vehicle-to-Everything (V2X)/etc.) and use-case (Handheld Computing, Wearables, IoT, M2M, etc.).

- Traffic Model (Applications, Service, Games, etc.)
    - Pattern of the data exchanged between a device/application/service and wireless network can be captured as mathematical models with specific requirements on throughput, latency, reliability, etc.
    - Alternatively, recorded traffic sessions for specific applications, games, scenarios can be used as traffic model.

- User Activity/Network Usage Model (Frequency, Idle/Inactivity/Activity)
    - Patterns of the user activity/inactivity, type and frequency of applications/ services used by the users. E.g., model can capture usage of a specific service (like video streaming) throughout the day, or specific days of the week, or any specific events (live streaming of sports/performing arts, etc.).
    - Pattern of upload/download of information from/to IoT/sensor device.

- Device Density Model
    - Device density models for various use-cases as defined in standards, or as per operator models.

- User Location/Mobility Model
    - Pattern of the user/device mobility.

The above information can either be standards (like 3GPP [16], IEEE, etc.) driven or data driven (captured by network infra and/or operators).

2.2.4 Modeling of End-to-End System

The digital twin for an end-to-end wireless system consists of twins for physical environment, radio environment, network deployment, service deployment, network subscriber population (i.e., user distribution model), and KPI/Event monitors corresponding to the objective of a ‘What-if’ analysis, as depicted in Figure 2-7. Although not explicitly shown, the RAN and CN models within the network model take into consideration modeling of transport network as well.

Individual models can be configured/enabled/disabled based on various factors, for example, the scenario under evaluation, specific use cases etc. These models can interact with each other at an aggregated level (as an example, for large scale scenarios) and/or can be following the standard 3GPP/O-RAN interfaces to interact with each other.

Depending on the problem (use-case) being solved, a subset of the components of the digital twin can be chosen as the focal point(s) for study and can be activated in plug-and-play mode. Output generated from the digital twin can contain various functional/performance metrics as well as network state/status information which can

be used for further analysis of the results, deriving inferences about the impact of configuration on network performance.

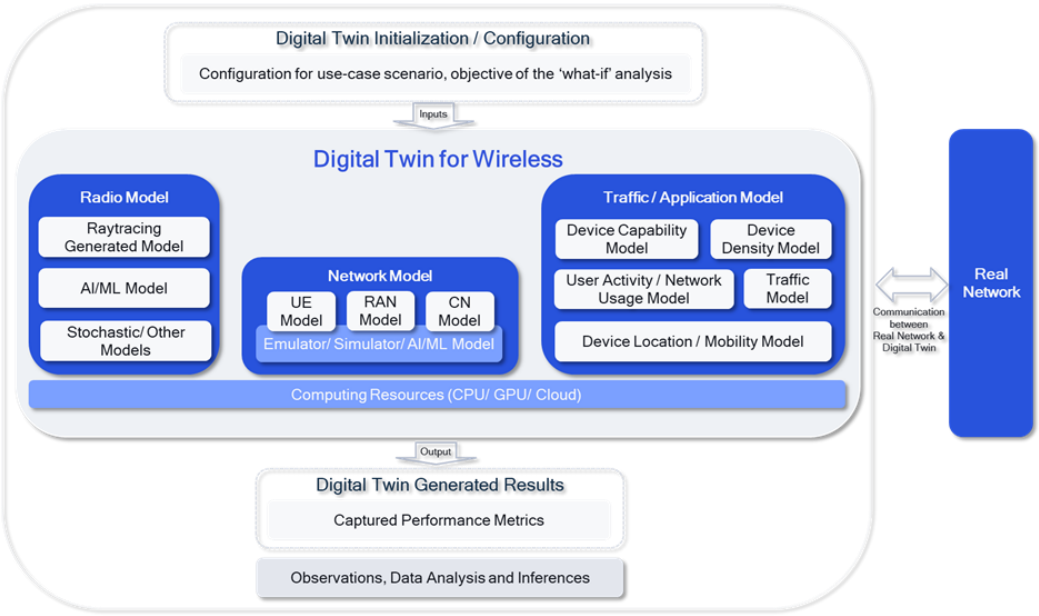


Figure 2-7: Modeling End-to-End System with Digital Twin

Enabling Technologies:

There are various techniques for modeling a DT-RAN. These techniques can be broadly categorized based on the modeling approach rather than being strictly divided into simulators, emulators, model-based, or data-driven categories, as these often overlap.

Simulation-based Techniques:

Simulation tools like Network Simulator (NS)-2, NS-3, and Objective Modular Network Testbed (OMNet++) employ detailed frameworks to mimic network behaviors. These simulators are highly accurate but do not operate in real-time and are often slow in performance estimation, which limits their efficiency in non-deterministic testing (NDT) scenarios [17].

Emulation-based Techniques:

Emulation techniques are aimed at executing the intended applications in a controlled communication environment and measuring the real underlying network behavior, where part of the communication architecture is implemented in a real setting [18]. However, network emulators suffer from inefficiencies in adapting to various network sizes and configurations.

Analytical Modeling Techniques:

Different from the simulators and emulators, analytical modeling techniques directly establish a mathematical relationship between the influencing factors and the performance metrics. Conventional analytical modeling methods (e.g., network calculus, Queuing theory etc.) have certain inherent limitations in modeling network

behavior [19]. While analytical modeling techniques are formalistic, they suffer from the following two major limitations: i) in complex network settings, such as ultra-highly dense heterogeneous networks, the model representing the system is mathematically intractable, and therefore a closed-form optimum or sub-optimum solution either does not exist or cannot be obtained, and ii) the use of simplifying assumptions might lead to inaccurate performance estimation, which makes the model-based techniques inflexible towards rapid and continuous network evolution.

Machine Learning-based Techniques:

Recently, driven by the advancement in the area of machine learning and abundant availability of data, data-driven techniques have been considered to model the network behavior. These techniques mainly involve learning a mapping function between the observations and the target output via a training process using real data, without

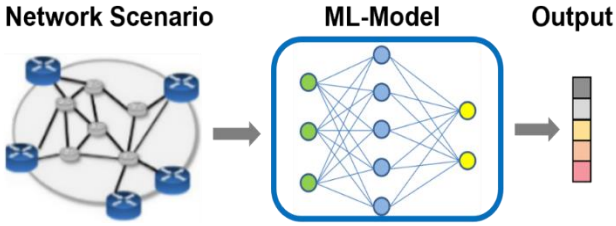


Figure 2-8: Data Driven Techniques, Input-Output Mapping

making any assumptions about the underlying network. This allows to build models with high accuracy by modeling the entire range of non-linear and multi-dimensional characteristics. Figure 2-8 depicts a generic schematic process of ML-based techniques. For example, the works in [20]-[22] have leveraged different flavors of neural networks, such as the Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) to extract information from network traffic and sequential dependencies of its parameters.

A downside of traditional ML-based techniques for modeling a DT-RAN is their reliance on large amounts of real data for training, which may not always be readily available or representative of all possible network conditions. Additionally, these models, such as CNNs and RNNs, can be computationally intensive and may require significant processing power, making them difficult to scale in real-time applications. They also tend to operate as black-box models, offering limited interpretability, which can make it challenging to understand or troubleshoot underlying network behavior, particularly in dynamic or unpredictable scenarios.

Real-Time and High Accuracy Modeling Techniques:

Many DT-RAN use cases, e.g. DT-RAN for AI/ML training, testing and performance assurance, network energy saving, site-specific network optimization and network automation require the DT-RAN models to be synchronized with the physical O-RAN network in real-time. Close to real-time simulations are needed for DT-RAN to catch up with the real network status. This include updating the RF propagation characteristics for each modeled UE according to real UE's moving positions, updating

radio link adaption parameters, traffic status, RRC states and call processes, etc. There is a trade-off between modeling accuracy and real-time performance of the DT-RAN. The accuracy and real-time requirement also vary with the use cases and the different scenarios that a DT-RAN is applied to. There is no *one-size-fits-all* modeling solution. Furthermore, for some applications requiring both real-time and high-accuracy, the model design and implementation become extremely challenging given the limited computing resource and cost budget for DT-RAN compared to the real network, which is one of the key technology gaps to be solved by the industry before DT-RAN concept can be realized in real-world.
Achieving real-time synchronization comes with significant costs, particularly in terms of computational demand and continuous updates. The level of real-time synchronization required depends on the specific application. Critical applications like network performance assurance or energy saving justify real-time updates due to their operational impact, while less demanding applications, such as long-term planning, can afford relaxed real-time constraints, making the trade-off between cost and performance more manageable.

Several unique features of DT-RAN make it possible to fill this technology gap if they are fully leveraged:

1. Real-time digital twin models are not truly real-time like the physical network. DT-RAN models can reflect the truth of the physical world at a certain instant of time when there is an application observing the DT-RAN from outside.
2. Digital twin model is not an all-encompassing mirror of everything happening in the physical network. It reflects only some selected aspects of the physical world that are of interest to other applications leveraging the DT model.
3. The accuracy and real-time requirement of DT-RAN depend on its applications.

Leveraging the above features of digital twin, potential end-to-end (E2E) modeling solutions can be explored. For example, as illustrated in Figure 2-9, digital twin models can be divided into two parts: 1) a non-real-time part including an offline RF grid generator and 2) a real-time part including the mobility and RF model, RAN function model and O-Cloud model incorporated into the O-RAN architecture.

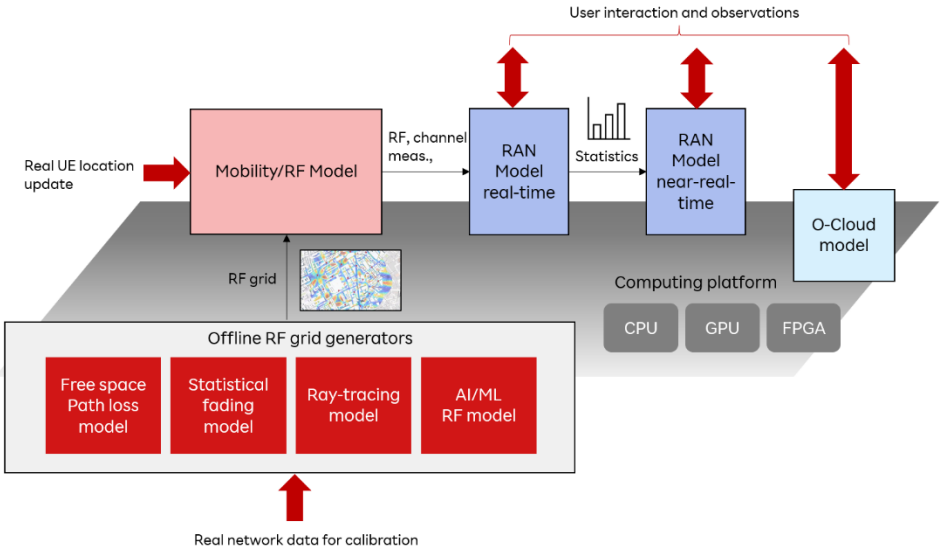


Figure 2-9: E2E Network Digital Twin Modeling

The offline RF grid generator can accurately model the large-scale RF propagation characteristics (e.g. signal strength, interference, angular spread etc.) at each position of an interested geographic area and create a RF grid map. To model the RF propagation environment accurately, compute-intensive technologies, e.g. ray tracing, can be used. The computation can be done offline which allows longer computing time with limited computing resource. The RF grid map can be uploaded to the mobility and RF model runtime. When an update of UE position is triggered by the UE mobility model, the RF propagation characteristics of the new UE location can be read directly from the RF grid map without recomputing large-scale characteristic in runtime. Some fast-varying small-scale RF characteristics can be modeled with statistic-based algorithm and added on top of the RF grid map. The high layer models can subsequently be updated with the latest channel measurements which trigger higher layer state changes. In some applications where RF modeling accuracy requirement can be relaxed, the number of rays to be traced in the ray-tracing model can be reduced, and/or some simpler RF models can be adopted, e.g. statistics-based RF modeling techniques or even free space path loss models, to minimize the complexity.

The RAN function model replicates protocol stack behaviors of the physical network, e.g. traffic scheduling, link adaption, beamforming, RLC retransmission, RRC configuration, mobility and handover handling etc. The model can be further divided into two parts- 1) The real-time part models the real-time, per-Transmission Time Interval (TTI) control loops in the network, e.g. traffic scheduling, link adaption, beamforming, RLC and Packet Data Ciphering Protocol (PDCP) etc., which normally happens on the user-plane, and 2) The near-real-time part models the higher layer control and management functions of the network, e.g. RRC configuration, mobility and handover management and Orchestration And Management (OAM) FCAPS etc. For the real-time RAN model which requires high computing resource, statistical algorithms which model the statistical KPI performance of the real network can be used instead of replicating all behavioral details of the real network in the DT-RAN at

a per TTI level. AI/ML technology can also be leveraged to learn the statistical behavior of a real network in correlation with various observed mobility and traffic conditions, and predict the statistics for a new condition in the DT-RAN. For the near-real-time RAN model, the computing resource requirement is relatively relaxed but accuracy is still critical. In this case, the digital twin can run the same software business logic as the real network to create a high-fidelity digital replica. Network function virtualization, containerization and user/control plane disaggregation in 5G network make it possible and easier to reuse the control and management plane network implementation in the DT-RAN to create a realistic replica and keep it synchronized with the real network.

Graph Neural Network (GNN)-based DT-RAN Modeling:

While traditional ML techniques (non-relational) have been successful in domains other than wireless network modeling (such as computer vision and natural language processing), they still fall short when it comes to modeling the intricate dynamics of wireless networks. This is primarily related to the fact that wireless networks usually have a lot of topological uncertainty. This can be attributed to a wide range of reasons including UE mobility, traffic demand, network heterogeneity, etc. Relying only on the traditional ML-based techniques to model such dynamics necessitates the collection of a large amount of data from real networks to ensure that the model is trained on several corner cases as well. Collecting such a vast amount of diverse data from real networks is extremely difficult. Therefore, the generalizability and scalability of such techniques are challenging in wireless network. Importantly, traditional ML techniques are developed under the assumption that the data is in Euclidean space. In wireless networks, however, the data is mainly in non-Euclidean space. Therefore, traditional ML techniques may not be a good fit for modeling of wireless networks, which motivates the consideration of graph-based techniques for modeling of network digital twins.

Graph Neural Networks (GNN) [17] leverage the underlying graph structure of the real networks to address the shortcomings associated with the traditional ML techniques. GNN-based techniques allow to store node-level hidden states and update them at each iteration. The update process considers the underlying dependencies of different elements of a network.

Connectivity in GNN is represented via an adjacency matrix and the prediction tasks that can be performed are divided into graph-level prediction tasks, node-level prediction tasks, and edge-level prediction tasks.

- In graph-level tasks, the goal is to predict the property of an entire graph, which is analogous to the process of labeling images, or sentiment analysis of a text.
- In node-level tasks, the objective is predicting the identity or role of each node within a graph, similar to the task of image segmentation or speech prediction.
- Finally, edge-level tasks perform predictions for the links of the GNN graph.

GNN-based techniques operate with permutation-invariant modeling principles implying they can handle different rotation variants of the same network topology.

Such capability allows working with data in non-Euclidean space, which represents wireless domain applications more accurately. Consequently, the model requires lesser data and parameters to train, adding to its generalizability and scalability in networking applications. Network heterogeneity is another challenge in traditional ML techniques that demands for a complex model, which in the case of GNN, is natively handled. This further adds to the flexibility of using GNN as a technique for network digital twin modeling.

Indeed, ML techniques like Multi-Layer Perceptron (MLP) and CNN are designed for image-processing tasks and are not tailored to wireless networks. In general, MLPs and CNNs offer good performance on small-scale networks but they suffer from poor generalizability and scalability. To improve the scalability and generalizability of DTN, the key idea in GNN is to incorporate the structure of the target task (e.g., prediction of link-quality) into the neural network architecture. The main benefit of GNNs as compared to other Deep NNs is the ability to generalize to different problem scales. For example, one can train the GNNs on small-scale problems and apply them to problems of larger scale. For example, for a beamforming problem, a GNN trained on a network with 50 users can achieve near-optimal performance in a larger network with 1000 users [24]. In general, the required number of training samples for GNNs is much smaller than that for MLPs.

In the literature, GNNs have already been proven as an effective technique for RAN use cases. In [25], a Deep Reinforcement Learning (DRL) solution is proposed that uses the underlying O-RAN architecture graph to learn the weights of the GNN for optimal user-cell association. In the paper, UEs are represented by nodes and the quality of the wireless links is given by the edge weights. GNN is used in [26] for modeling radio propagation in wireless networks. In their approach, the GNN nodes correspond to locations and the edges represent spatial and ray-tracing relationships between these locations. [27] uses a generalizable GNN-based technique to solve the problem of 5G RAN/MEC slicing and admission control in metropolitan networks. The authors claim that the GNN approach converges faster than comparable DRL methods.

The aforementioned benefits and the existing literature make GNNs a key enabler for modeling DT tasks in RAN. GNNs as enabler of DTN is also reflected in the latest ITU AI/ML challenge which focuses on building DT based on GNNs using dataset from a real network. However, despite the empirical successes, the design guidelines still remain vague, which may hinder the practical implementation of GNN based DTN. Hence, more research in this area is encouraged.

GNNs have become increasingly crucial for modeling complex network structures, such as in the creation of DT-RAN. The effectiveness of GNNs hinges on a suite of enabling technologies and methodologies that ensure efficient data handling, computation, deployment, and algorithms efficiency [17]-[19].

Data handling:

Graph databases are integral to GNN modeling, offering an efficient way to manage graph-structured data. Their design is specifically aimed at meeting the needs of

network data, which is inherently graph-based, making them ideal for GNN applications. These databases excel in representing data as nodes and edges, reflecting how GNNs process information. This capability is crucial for modeling complex networks, where relationships are as significant as the entities. Graph databases handle intricate and dynamic network relationships well, efficiently managing complex interconnections found in various real-world networks such as social, biological, and telecommunication systems [20].

These databases boast specialized query languages and excel in traversal performance, enabling efficient data retrieval and analysis of interconnected data. This feature is particularly useful in applications requiring rapid network traversal, like real-time recommendation systems. The data handling methods of graph databases align closely with GNN model requirements, allowing straightforward and efficient data feeding without extensive restructuring. They can dynamically update network data, which is crucial for applications like fraud detection that rely on real-time data. Graph databases scale effectively with network size. Moreover, they are optimized for operations commonly performed in GNNs, enhancing data retrieval and pre-processing efficiency. Graph databases are vital in GNN modeling, offering an optimized, efficient system for managing graph-structured data, which is crucial for effective and accurate network modeling.

Graph Data Management/Processing is essential in preparing data for GNN modeling, involving multiple steps to convert raw data into a structured, usable format for accurate network modeling. This process includes 1) data cleaning to eliminate inaccuracies and irrelevant information, and 2) data transformation, where data is represented as nodes and edges to reflect entities and their relationships. A significant challenge is to accurately map real-world relationships into a graph structure. Normalization is another critical step, standardizing variables to ensure no single feature disproportionately affects the model behavior. Data integration is complex yet vital, combining data from varied sources into a unified graph structure. These steps collectively assure the quality of the data, ensuring it is clean, structured, normalized, and integrated. The accuracy of data representation in graph format is critical for the GNN's ability to model network characteristics correctly, as any misrepresentation can lead to inaccurate predictions.

Computation:

Computation efficiency is pivotal in GNNs, given their application in modeling complex and often large-scale networks. Enhancing this efficiency centres around minimizing computational demands and accelerating training and inference processes. One fundamental approach is parallel processing, where computational tasks are distributed across multiple processors or cores. In GNNs, this entails distributed training across several machines or Graphics Processing Units (GPUs), beneficial for extensive graph datasets where training on a single machine would be inefficient. Additionally, batch processing allows for simultaneous handling of multiple data samples, significantly speeding up both training and inference phases [21].

Another crucial aspect is GPU acceleration, which aligns well with GNNs due to its inherently parallelizable nature. GPUs, with their multitude of efficient cores, are adept at handling the matrix and vector operations common in GNN computations. They excel in performing fast matrix operations, a frequent requirement in GNNs, and their substantial memory bandwidth is vital for managing large datasets and models. Alongside hardware acceleration, optimizing GNN architecture is also a key. This includes designing lightweight yet effective models, optimizing individual layers (such as using sparse rather than dense matrix operations), and integrating advanced algorithms that reduce computational complexity without sacrificing learning capacity. Other strategies like model quantization, which reduces the precision of parameters, and model pruning, which eliminates non-essential model components, further streamline computation. These comprehensive approaches in hardware utilization, architectural refinement, and algorithmic innovation collectively empower GNNs to efficiently process complex, real-world network data, broadening their applicability across various domains.

Deployment:

The deployment strategy for GNN models is vital, as it greatly influences their performance and suitability for various applications. The decision to use cloud computing, edge computing, or a hybrid of both hinges on factors like network system requirements, data characteristics, and desired outcomes. Cloud computing offers substantial computational resources, storage, and scalability, making it ideal for complex GNN models and large data volumes. It provides high scalability, centralized management, and advanced analytics tools. However, its limitation with latency can be a drawback for real-time data processing needs.

Edge computing, on the other hand, processes data close to its source, such as IoT devices, which is beneficial for GNN models requiring immediate insights or operating under limited connectivity. This approach minimizes latency, supports real-time processing, and optimizes bandwidth usage by reducing data transfer to the cloud. The trade-off here is the relatively low computational power and storage capacity compared to cloud platforms. The hybrid approach merges the strengths of both cloud and edge computing, allowing data processing both locally and in the cloud. It combines edge computing's speed with the cloud's power, offering flexibility and scalability. This method also enhances security and reliability by distributing processing. However, the hybrid model's complexity in implementation and management requires careful planning and coordination for efficient operation.

Algorithm efficiency:

The efficiency of algorithms in GNN modeling is crucial for handling complex networks effectively and practically. This efficiency has been significantly boosted by the advent of open-source libraries, which provide pre-built and optimized algorithms for GNNs. Libraries like PyTorch Geometric, Deep Graph Library (DGL), and TensorFlow Graphics offer a suite of tools and algorithms, including optimized implementations of standard GNN architectures like Graph Convolutional Networks (GCNs) and Graph Attention Networks (GATs). These resources enable developers to avoid the intricate

process of coding these models from scratch, leading to a faster and more reliable development cycle [23].

Various techniques further enhance the efficiency of GNNs. Graph sampling methods like node, layer, and subgraph sampling reduce computational complexity, making training on large networks more manageable. Parallel processing, leveraging modern GPUs and Tensor Processing Units (TPUs), expedites data processing and model training. Optimizing message-passing algorithms, a core component of GNNs, also boosts performance by minimizing computational overhead. Additionally, model pruning, and compression techniques reduce the size of GNN models, maintaining performance while cutting down on computation and memory usage. Transfer learning, using pre-trained models on similar tasks, saves resources and reduces the computational burden by requiring minimal additional training.

Efficiency in GNN modeling is also achieved through architecture optimization, customizing network structures to align with specific data characteristics and tasks. This includes tailoring layers, activation functions, or the entire network design, leading to more efficient models. By integrating open-source libraries with these efficiency-enhancing techniques, GNN modeling becomes not only more efficient but also more adaptable to various applications. This combination enables the creation of powerful, computationally manageable GNN models, adept at tackling the complexities of network data analysis and prediction.

In summary, to enable GNN modeling for DT, the integration of data handling, computational efficiency, deployment strategies, and algorithmic improvements is crucial. Graph databases ensure GNNs receive accurate, structured data, while computational efficiency is boosted through techniques like parallel processing and GPU acceleration. Deployment strategies are tailored to specific needs, balancing power, and speed. Algorithmic efficiency, enhanced by open-source libraries and optimization methods, streamlines GNN modeling. Together, these elements ensure GNN models are not only powerful but also adaptable, crucial for creating effective DTNs for a wide range of network analysis and prediction tasks.

Example Use Case:

Use Case 1: GNN for Network Energy Efficiency in DT-RAN

With the rise of AI/ML approaches for optimization, deriving strategies that enhance network operations without disrupting real-time processes is becoming increasingly important. In this context, DT-RAN, a digital replica of physical RAN, serves as a powerful tool for testing and validating various strategies before they are applied to real networks. DT-RAN allows for planning and forecasting the resource usage to identify under-utilized network resources that can be shut down or re-purposed to reduce energy consumption. DT-RAN can potentially help in achieving substantial network energy saving driven by a combination of environmental, economic, and operational factors. It can model the impact of various network energy saving strategies on the network performance using a virtual representation of the network topology, dynamics and operational behavior to arrive at optimal models and

thresholds without compromising the end-user QoS. However, maintaining the right balance between end-user's QoS and network energy saving is quite challenging and intractable from an analytical perspective, in a dynamic, heterogeneous, and large-scale RAN.

To address such challenges, data-driven techniques such as MLPs and CNNs have been considered. These techniques, while quite successful in other domains,

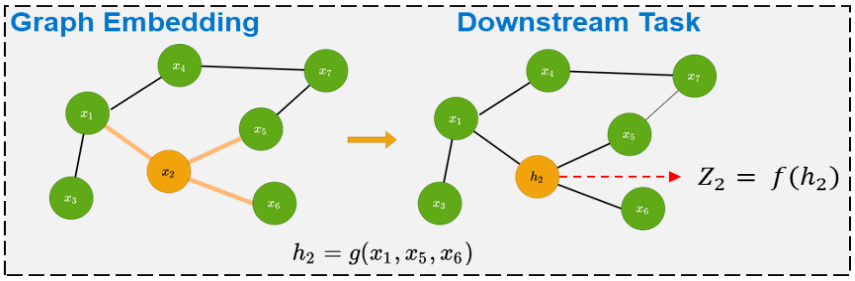


Figure 2-10: Graph Embedding using GNN

however, can only address either localized issues or apply to small-scale wireless networks. There is, therefore, a need for methods that can inherently exploit the wireless network topologies and can scale and be generalizable for large cellular network resource management aspects.

As shown in Figure 2-10, the key principle in GNN is graph embedding, which involves taking each node and its neighbours' features and encoding them into a vector representation (embedding) using a permutationally invariant process. This means that the order of nodes in the graph does not affect the embedding, ensuring that the resulting vectors accurately capture the relationships and features of each node within the graph. In RAN use cases, the nodes are the network elements, such as user equipment, base stations, etc. and the edges are the communication links among them. Once the node embeddings are obtained, they can be used in downstream tasks in xApps and rApps for specific network optimization goals such as reduction of energy consumption while maintaining the target QoS requirements. By analyzing these embeddings, network operators can make informed decisions to optimize resource allocation, improve network performance, and enhance overall efficiency. Additionally, these embeddings facilitate more accurate predictive modeling and anomaly detection, further enabling proactive management and optimization of the network.

As shown in Figure 2-11, a trained GNN is leveraged to model the relationship among the network elements in DT-RAN. The goal is to use GNNs to model the behavior of the network elements for any changes in the network and evaluate their impact on the overall network energy consumption and QoS without risking degradation of the real network. The DT-RAN as a virtual replica of RAN, receives real-time network data. Such data is composed of the current state of the network both at the component-level and at the network-level. The DT-RAN constructs the network's graph using that data. The network graph can be used by network operators to evaluate various 'what-if' scenarios for actuation of the Network Energy Saving (NES) strategies and to

ascertain their impact on overall network energy consumption and target QoS requirement. Such ‘what-if’ scenarios include changes in network topology, UE traffic demands, routing, scheduling, UE association, admission control, handover, etc. For instance, by simulating these scenarios, operators can predict how different configurations affect energy usage. This approach helps identify opportunities for optimizing network operations, such as reducing redundant transmissions, minimizing idle times, and adjusting resource allocation dynamically. Through effective analysis of these scenarios, operators can make informed decisions that reduce energy consumption while maintaining or even enhancing network performance, leading to significant energy savings. The network graph at this stage, can be as complex as the actual RAN, reflecting the intricate interactions between network components. Furthermore, these scenarios can vary in complexity, allowing for different levels of granularity within the network graph to be explored.

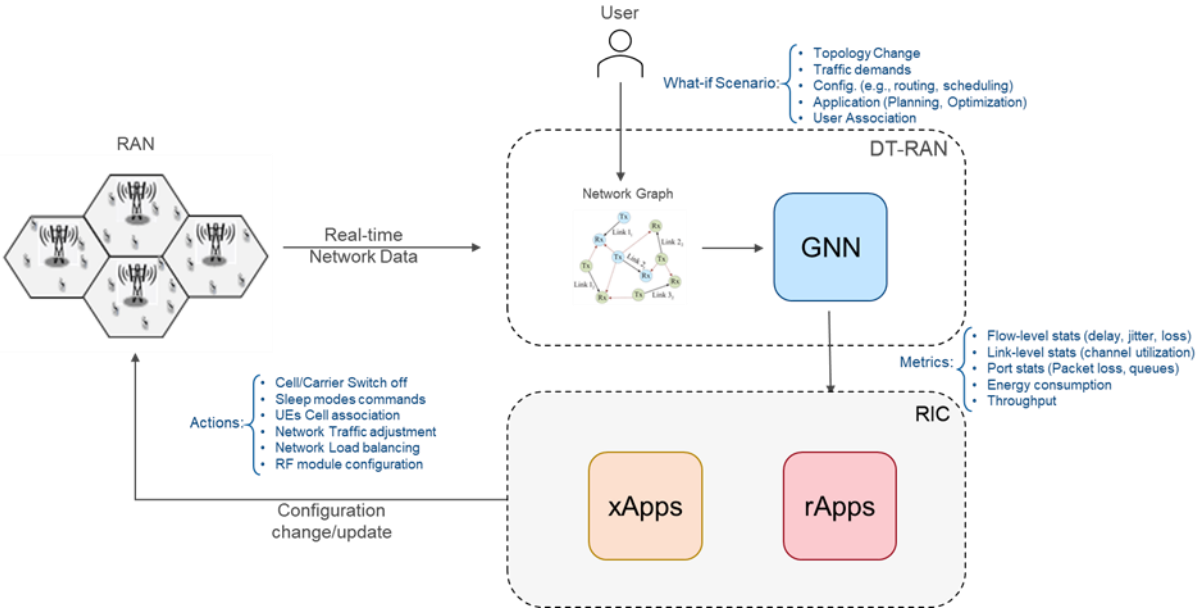


Figure 2-11: GNN for Network Energy Efficiency in DT-RAN

The GNN processes the network configuration incorporating the 'what-if' scenarios and produces network energy consumption data, along with other QoS metrics, including network flow-level statistics, link-level statistics, and network throughput. This information can be utilized in both near-real-time applications (xApps) and non-real-time applications (rApps) deployed on the RIC, following the O-RAN architecture. Alternatively, if DTN techniques are employed, the GNN model can be directly embedded within xApps or rApps. This integration allows xApps and rApps in the RIC to issue informed control commands to the RAN, ensuring that the target QoS is maintained. These control commands may include cell/carrier switch off, sleep modes commands, network load balancing, as well as RF module configuration.

2.2.5 Radio Spectrum Awareness and Emitter Activity Modeling

As useful radio spectrum is an incredibly scarce resource that must be re-used and shared efficiently within and between networks in many cases, awareness of emissions and spectrum activity is becoming increasingly important to enable efficient intra-network operation and optimization as well as inter-network and inter-technology coordination and optimization. To this end, DT-RAN should optimally employ efficient spectrum awareness technologies distributed across various network elements in order to enable intelligent monitoring, adaptation, and reaction to a broad range of spectrum activity in order to facilitate better use of the spectrum through increased spectral efficiency and spatial re-use. By detecting emitters, and localizing within the DT-RAN model, we can track where emitters are, and what interference they may be generating from which propagation paths. This could be used to help with training or testing network functions such as scheduling, beamforming, beam nulling etc. to work well in the presence of other emitters, Electromagnetic Interference (EMI), or objects in space which are contending for the spectrum. In addition to improving spectral efficiency, this can help to automate operations, streamline OAM and fault diagnostics within networks, mitigate security threats and interference, and help to enable a deeper awareness of activity in the physical world in ISAC related next generation use cases. By tracking RF emitters, power levels, MIMO channel statistics, localization estimates, access times statistics, and other statistics within the DT-RAN from one or more network elements, a rich set of models can be built to help represent the behavior of the interferers, nearby networks, and nearby activity. This can be used by xApps and other services to help optimize how spectrum is allocated, how beams, nulls, and frequency sub-bands are scheduled, and to help avoid or operate through network degradation or known sources of interference or adjacent networks. Furthermore, this may be a fundamental enabler of future shared-spectrum band access strategies, helping to inform future Citizen Broadband Radio Service Spectrum Access System (CBRS SAS) like sharing orchestrators with rich and detailed information about spectrum access and propagation, leading to better and more efficient heterogeneous re-use in complex and dense environments.

Figure 2-12: RAN Digital Twin Localization of Non-Network Emitters, Arlington

In one example shown in Figure 2-12, we illustrate how gNBs can help to detect and localize sources of EMI and unauthorized or out-of-network gNBs which are co-existing in an FR1 spectrum band. Here, knowledge of these devices from sensing events from each sector can be used and combined within the RIC based DT in order to help localize and trigger detailed alerts to operators, and to enable base stations to track the interferers in order to help simulate and mitigate interference through spatial processing, cancellation, and resource allocation.

Enabling Technologies:

Machine learning, and specifically deep learning have completely transformed what is possible in the realm of spectrum sensing. Leveraging similar techniques to those of object detection and tracking in the visual domain (e.g., which have enabled self-driving cars, drones, and numerous forms of visual autonomy), many of these same technologies can be adapted to work on complex valued radio signals and spectrum to help enable very low cost, low power, low latency, and highly adaptable data-driven sensing for a wide range of types of emitters and phenomena with very high accuracy and sensitivity. Figure 2-13 below illustrates one such software capability, passively detecting various 4G and 5G emitters in band, along with channel statistics from short sub-millisecond observations to help inform DT-RAN RF Emitter maps using compact and descriptive structured data, which can be efficiently and compactly passed over interfaces to xApps, rApps, or other network elements.

As these machine learning models for sensing are natively data-driven, they can be rapidly retrained on many types of communication emitters including cellular, IoT, and Wireless Local Area Networks (WLAN) type emitters as well as non-communications EMI, radar, and wireless phenomena. This allows for rapid out-of-distribution or anomaly detection, curation, and retraining on new data and models and deployment of new sensing capabilities to the edge at network elements or RIC applications, to

improve spectrum awareness under new conditions. As embedded and mobile silicon is increasingly employing neural network acceleration hardware such as Neural Processing Units (NPUs), Data Processing Units (DPUs), AI-Engines, GPUs, Matrix extensions, etc., these models are increasingly able to be deployed to edge network elements such as O-RU or O-DU elements with minimal cost, operation, and network impact or cost to vendors or operators, making this increasingly appealing for deployment.

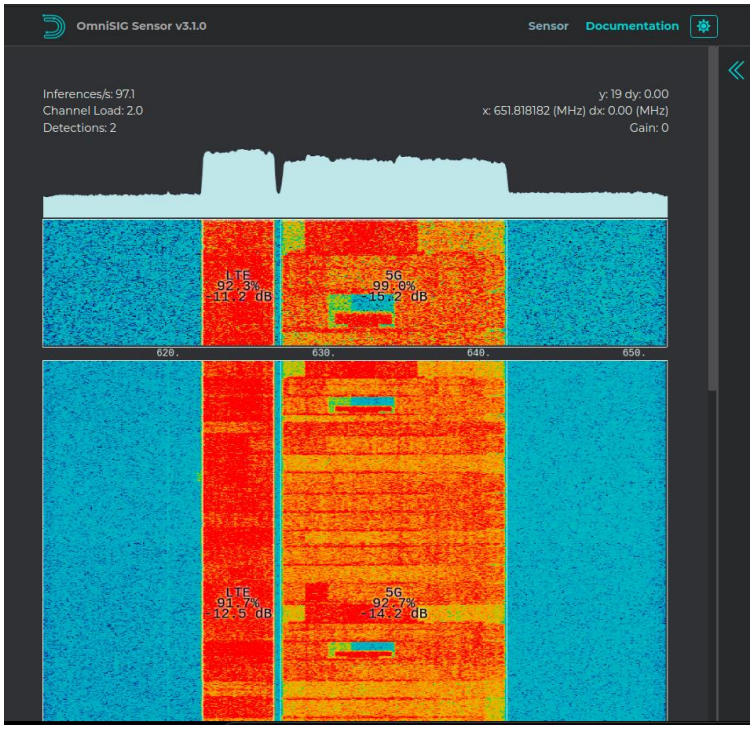


Figure 2-13: Wireless Spectrum Awareness Software Running over the Air

Example Use Case:

Use Case 1: DT-RAN based Spectrum Awareness

Numerous types of nearby RF emitters can impact RAN performance and are key use cases to be enabled by deploying DT-RAN based spectrum awareness. These emissions include:

- Detecting electromagnetic interference, jamming, or other emitters that may cause degradation in RAN performance or disruptions in coverage, requiring mitigation, avoidance, or enforcement actions to clear spectrum.
- Detecting and monitoring unauthorized devices, physical layer attacks, or other types of security threats which may be operating on the air interface.
- Enabling future shared spectrum bands, by helping to identify usage and signal strength and channel statistics of adjacent networks, helping to orchestrate more efficient and dense re-use without degrading performance.
- Enabling efficient operating in unlicensed bands such as for New Radio-Unlicensed (NR-U), Private 5G deployments or related future technologies by identifying

emitters of various technologies, spatial and power statistics of emitters, and by helping to coordinate and mitigate interference between elements.

For many of these use cases, the DT-RAN may store representations of the events, emitter information, location, power level, and spatial statistics of emission events, and may use these in combination with various channel or propagation models including those discussed in Section 2.2.2 to help inform a decision about spatial processing, scheduling, band assignment, etc. Numerous automated responses or reactions can then be deployed within the RAN such as:

- Spatial reactivity: Reacting to trigger additional spatial processing to reduce interference or effects on the RAN such as by steering beams, nulls, resources, or other effects.
- Processing reactivity: Leveraging additional processing stages to mitigate interference, for instance by further training receiver components in the presence of various types of noise or interference to improve performance.
- Enabling feedback: Actively communicating between network elements within the network, between networks, or between network orchestrators such as future SAS services in order to help reduce interference in certain bands, spatial modes, channel access modes, power levels, or antenna configurations.
- Analytics: Enabling new use cases through real time information and analytics of RF emitter behavior and movement in the physical world, emitter locations or behaviors, or reflector behaviors or movements which may be extracted through processing of channel statistics.

2.3 Computing

As RANs are evolving towards increased heterogeneity, dynamicity, and complexity, it is natural that the DT-RANs require advanced technical capabilities to faithfully replicate the physical network traits in the virtual domain [28]. To enable seamless synchronization between a DT-RAN and its physical counterpart, fast and efficient processing capability is essential, which calls for accelerated computing that ensures a proactive response of digital twins to any change or event occurring in the underlying real networks.

2.3.1 Accelerated Computing

Accelerated computing is a specific style of computing, where data-intensive part of an application is processed on a specialized acceleration device that can speed up the processing capability, usually utilizing the parallel processing capability of several tasks simultaneously, instead of in a linear or serial fashion as is usually done in traditional general-purpose processing. As mentioned in the previous section, DT-RANs are poised to harness the power of AI/ML techniques for enhanced decision making and predictive analysis capabilities. Accelerated computing is crucial in aiding

faster execution of training and inferencing of AI/ML models running in DT-RANs [29]. 6G networks will be dynamic, and it is anticipated that online training for AI/ML models will be crucial. One example is Deep Reinforcement Learning based models, where a neural network embedded in the DRL agent undergoes continuous updates in response to the dynamic environment. Latency is important here and hence the training and inferencing speeds are crucial. Accelerated computing is essential in ascertaining the demanding training/inferencing time for these DT-RAN models [5]. The emerging application landscape of 6G is diverse, proliferated by new technology enablers like integrating communication with LiDAR/RADAR sensing, RIS utilizing enormous antenna arrays and advanced beamforming techniques and Terahertz (THz) frequency communication. Supporting physically accurate radio wave propagation modeling in DT-RAN to mimic the network behavior for these wide range of technologies with high-fidelity requires sophisticated tools like ray tracing, as explained in Section 2.2.2. Ray tracing at-scale is exceedingly computationally intensive, and accelerated computing is vital in supporting ray tracing-based channel modeling in DT-RAN [5].

It is evident that DT-RANs will be fueled by massive amount of data, either collected from various physical sources (like sensors, IoT devices and network elements), or generated through simulation [30], as explained in Section 2.1. The handling and processing of this large volume of data and deriving analytical insights for live network entails accelerated computing as an essential tool for DT-RAN. Time-sensitive use cases like anomaly/fault detection require near-instantaneous response time from DT-RAN to the underlying physical network. Advanced computing capabilities contribute to minimizing latency in response time, enabling DT-RAN to provide prompt and timely feedback to the physical networks and alleviate the risks of network downtime. The modularity of digital twins representing a physical network can vary depending on the underlying network deployment topology. For example, instead of being centralized, DT-RAN can be distributed across multiple edge devices and cloud infrastructure. Accelerated computing aids in efficient distributed computing, enabling digital twin models to seamlessly interact with various parts of the network in real-time and share feedback across the network [31]. As the emerging technologies are focusing more and more on sustainability, there is no doubt that the approach towards green communication will shape the deployment landscape for 6G networks. At that juncture, energy efficient computing is paramount to the proliferation of digital twin technology in the telecommunication domain [32]. Accelerated computing technologies can help optimizing energy consumption of digital twins and achieving the goal of minimizing digital twin’s carbon footprint.

To summarize, the combination of increased complexity, real-time requirements, massive data processing, integration of AI, distributed computing, low latency demands, highly scalable modeling and energy efficiency considerations cumulatively contribute to the need for accelerated computing in digital twins for the next generation RAN.

Enabling Technologies:

Technologies enabling accelerated computing in DT-RANs can be broadly categorized into two buckets – 1) hardware-centric tools and 2) software-centric tools. Optimized handshaking between these two brings in the best-in-class performance and enhanced efficiency in the twin domain. Following are the set of accelerated computing tools that can improve the efficiency and performance of DT-RANs.

Acceleration Hardware and Software:

Programmable hardware accelerators like GPUs, with their massive parallel processing capability are highly efficient in accelerating real-time simulation, analytics, and processing of AI/ML workloads. Tensor core enabled GPUs are specifically designed to offer mixed-precision computing and in turn, acceleration of AI/ML model training and inferencing. With GPU acceleration, the development and deployment of AI/ML models in DT-RAN can be significantly faster. GPUs excel at handling data-intensive tasks, making them suitable for analyzing and deriving meaningful insights from massive and complex datasets.

Software programming environments (including Application Programming Interfaces (APIs) and programming frameworks) designed for parallel processing can simplify the development of GPU based DT-RAN models. Examples include Compute Unified Device Architecture (CUDA), Open Computing Language (OpenCL), Open Accelerators (OpenACC) and Open Multi-Processing (OpenMP). These tools help AI/ML algorithm developers/programmers to exploit parallelism in their code and efficiently utilize the compute capabilities of software-defined accelerators. GPU-accelerated software libraries facilitate DT-RAN model development that can take full advantage of accelerated computing resources. Examples of software libraries optimized for AI/ML based models and algorithms development include CUDA Deep Neural Network (cuDNN), DeepStream Software Development Kit (SDK), TensorRT, etc. Alongside, optimized software libraries for linear algebra (including CUDA Fast Fourier Transform (cuFFT), ArrayFire, Matrix Algebra for GPU Multi-core Architecture (MAGMA), CUDA Math library, and CUDA Basic Linear Algebra Subroutine (cuBLAS), among others) can provide further computation acceleration involving parallelizable mathematical operations associated with various algorithms running during the complex simulations of DT-RAN.

While programmable hardware accelerators aid acceleration of versatile workload processing in DT-RAN, specialized hardware accelerators that are custom-made for specific workload acceleration can be useful for targeted tasks in DT-RAN. As one example, Deep Learning Accelerator (DLA), a fixed-function inference accelerator tailor-made for processing neural network workloads can be used for accelerating deep learning operations in DT-RAN. Other examples of AI accelerators include Tensor Processing Unit (TPU), Field Programmable Gate Array (FPGA), Application-Specific Integrated Circuit (ASIC) and System-on-Chip (SoC). High Level Synthesis (HLS) tools enable design, optimization, and implementation of these custom hardware accelerators.

Multi-GPU Scaling and High-performance Compute:

System level simulation is an integral part of DT-RAN, and its compute requirement can vary over a wide range of scales – all the way from the simulation for a specific network site to the massive, city-scale simulation. Compute scalability, therefore, is an essential trait for DT-RAN. Leveraging the principle of parallelism in GPU, various DT-RAN workloads (e.g., data processing, analytics, simulation, complex model training

and inferencing) can be parallelized and distributed across multiple GPUs simultaneously – an acceleration approach known as multi-GPU scaling. Utilizing multi-GPU scaling can significantly increase the processing capability of DT-RAN and can be beneficial for processing large datasets and intricate simulations at a faster pace. For DT-RAN modeling and simulation at city-scale, High Performance Computing (HPC) systems with clusters of powerful processors and accelerators (including GPUs) can provide exceedingly high-speed processing capability and enormous computational power needed for highly complex modeling, real-time optimization, and Faster-Than-Real-Time (FTRT) simulation.

Cloud Compute:

DT-RAN operates in a dynamic environment and its compute need varies over time. While scalability is one important aspect of accelerated computing in DT-RAN, the other essential trait is the flexibility in provisioning accelerated computing resources on an ‘as-needed’ basis, to make the DT-RAN not only compute efficient, but also resource and energy efficient. DT-RAN deployed on a cloud platform offers that flexibility and scalability, by allowing the accelerated computing resources such as GPU instances to be provisioned ‘on-demand’ and thereby, offering agile and dynamic scaling of computational capacity based on the requirements of DT-RAN workload.

Distributed/Edge Compute:

For mission critical applications with stringent latency requirements, DT-RAN’s response time needs to be of low latency. Enabling centralized accelerated computing for DT-RAN (for example, at near edge) may not be adequate in meeting the time budget for latency-critical simulations. Edge computing alleviates this issue by bringing computation resources closer to the data source, thereby reducing latency in data transfer, and enabling real-time processing. Deploying accelerated computing tools (e.g., hardware accelerators) closer to the edge enhances both compute speed and low-latency responsiveness of DT-RAN.

Example Use Cases:

Use Case 1: DT-RAN for Site-specific Network Planning and Optimization

As mentioned before, ray tracing is an integral part of DT-RAN for enabling physically accurate model of radio wave propagation environment, which is essential for site specific network optimization. Making at-scale RF ray tracing a reality requires programmable hardware accelerators supporting the essential mathematical operations acceleration for ray tracing and optimized software libraries for implementing the ray-tracing pipeline. As one example, application frameworks like OptiX can be used to exploit the accelerated computing offered by RTX GPUs and achieve optimum ray tracing performance for DT-RAN RF environment modeling [5].

Use Case 2: DT-RAN for Network Automation

An important aspect of network automation is enabling zero-touch networks, which is capable of self-healing and adjustments based on autonomous analysis of network data and activity. Predictive analysis and anomaly/fault detection are essential traits

of zero-touch networks. Accelerated computing empowered DT-RAN can offer fast and efficient simulation, root-cause analysis, and fault-correcting actions with low-latency response time, enabling on-time auto-remediation to the anomalies of the underlying physical network.

2.4 Visualization & Trustworthiness Management

Visualization is the bridge between the internal complexities of a DT-RAN and the user. It enables users to interact with the digital twin model and understand its outcomes. Visualization also aids as one of the important tools in monitoring the DT-RAN and ensures its trustworthiness over long run.

2.4.1 Visualization

A typical DT-RAN will generate an enormous volume of large-scale network data, and this data will be characterized by its size, dimension, and heterogeneity [28]. The industry has always relied upon simple tools such as interactive dashboards and charts to visualize the operations of the RAN. This is now complemented, for the DT-RAN, by a new set of high-end graphics such as 3D models and Augmented Reality / Virtual Reality (AR/VR). High quality visualization brings the DT-RAN to life, enabling users to understand, monitor, analyze and interact with the underlying complexities of the DT-RAN models. This is necessary to support decision making, collaboration among co-workers, predictive analysis, training, optimization, and maintenance.

Enabling Technologies:

Broadly speaking, visualization tools and technologies for the DT-RAN can be split into five main groups. These can be used individually, or more commonly, combined to provide a visual interface for the DT-RAN. When used in a real RAN, these tools visualize the actual performance data from the RAN and helps the operator to monitor or manage the performance of the network. For the DT-RAN, the tools visualize the synthetic performance data of the DT-RAN to help the user to understand the underlying performance of the digital twin.

Performance Metrics Dashboards:
Performance metrics dashboards typically provide discrete information about the status or performance of different parts of the DT-RAN such as latency, throughput, packet loss, and jitter [33]. These dashboards display real-time and historical data from network elements for all or selected parts of the DT-RAN in a visually interactive format. Dashboards are primarily used for performance monitoring by providing at-a-glance information, e.g., to keep track of wireless channel conditions above a given threshold. They are also used for fault/alarm monitoring with display alerts, notifications, and alarms, e.g., to alert when wireless channel conditions deviate from acceptable levels. Dashboards help users of the DT-RAN to quickly identify and respond to network issues, failures, or security breaches, minimizing downtime and service disruptions.

Graphs, Charts and Heatmaps:

These provide an enhanced visualization of continuous information from the DT-RAN, enabling the user to visualize network traffic patterns, congestion, and bandwidth utilization across different parts of the network [34]. Graph/charts show trends, patterns, and relationships within the data and can be used to convey different types of information such as temporal trends, distributions, and correlations. Heatmaps use colour coded charts to represent the intensity or density of data across a spatial or temporal domain. An example is the traffic heatmap which shows areas of high or low traffic intensity in a DT-RAN environment, helping users identify bottlenecks, optimize routing, and allocate resources more effectively.

Network Topology and Flow Analysis Maps:

Network topology and flow analysis maps show the layout and structure of the components of the network and how data flows through these components [34]. The topology map is a visual representation of the network showing the components (e.g., cell site, user devices), connections, and relationships between components of the DT-RAN. The flow analysis map visualizes network traffic flows and traces the path of data packets as they traverse the network. By tracing the path taken by data packets from source to destination (including cell handovers), the user can understand the route for data packets, highlighting intermediate hops, latency, and potential points of congestion.

Geospatial Visualization:

Geospatial visualization techniques use realistic or virtual digital maps to represent spatial data and linkages in the DT-RAN [35]. They overlay network data onto geographic maps or satellite imagery, providing spatial context for network infrastructure deployed across different locations. Maps include GIS (Geographic Information System) overlays, satellite imagery, or 3D terrain models for particular geographic terrain where the DT-RAN which is being simulated is located. Maps can be combined with other visualization techniques, such as charts and graphs, to create predictive simulations of future scenarios, what-if analysis, or interactive simulations that allow users to control inputs and observe the outcomes for a DT-RAN [38]. Geospatial visualization is particularly useful for distributed networks, such as DT-RAN modules spanning across multiple sites.

High-end Graphics: 3D Models and AR/VR:

High-end graphics such as 3D models and AR/VR enable immersive visualization experiences for exploring and interacting with the DT-RAN [36][37]. 3D models use computer graphics to create a visual three-dimensional representation of the DT-RAN, and the spatial context for its operations. Many off-the-shelf visualization tools for digital twins use 3D models to create a virtual replica of the physical object or system. With AR, network data can be overlaid onto the physical world while with VR, a fully immersive virtual world environment is created for the DT-RAN to simulate network

behaviors, protocols, and scenarios in real-time or offline mode [36]. 3D, AR and VR models can be built with the Universal Scene Descriptor (USD), an open and extensible framework for describing, composing, simulating, and collaborating on 3D computer graphics data [39]. High-end graphics visualization is suitable for training, design reviews, or collaborative decision-making. For the DT-RAN, users can visualize the effects of changes to network configurations, traffic patterns, or routing protocols and assess their impact on network performance and reliability.

Example Use Case:

Use Case 1: DT-RAN for Network Performance Prediction

In a DT-RAN system-level simulator that scales to hundreds of base stations and tens of thousands of mobile users, visualization provides the user interface to interact with the DT-RAN to benchmark network efficiency, perform 'what-if' analysis under realistic conditions, and better understand the outcome of DT-RAN's simulation and evaluation. By running features, performance, and full-scale network testing early on, rather than waiting until deployment to see the impact of key design decisions, the DT-RAN dramatically reduces development cost and time to market. The ability to predict and visualize network performance under various conditions not only helps with network planning but also with operations of deployed networks, effectively reducing network downtime by figuring out 'early on' potential misconfigurations or network anomalies (including security threats) through easily interpretable DT-RAN's evaluation before these aberrations impact the live network.

2.4.2 Trustworthiness Management

A DT-RAN should be highly reliable to achieve trustworthy virtual-real interaction, including accurate reflection and reliable control of the physical network. Trustworthiness is at the core of DT-RAN's fundamental characteristics, which refers to the degree to which the virtual representation of the physical network can be relied upon to accurately and consistently reflect the real-world conditions and behaviors of the physical network. However, due to limited available network resources, and limited computing resources or degraded model performance, the trustworthiness of DT-RAN might degrade. For example, limited data transmission bandwidth and high network load can prevent the network from supporting real-time data collection for model construction and synchronous mapping of physical network. So, the results/outputs of the DT-RAN at the corresponding moment would not be trustworthy, and we need to monitor the trustworthiness of the DT-RAN to make necessary adjustments when there is a decline in DT-RAN's trustworthiness.

Enabling Technologies:

A DT-RAN is constructed by integrating multiple models. The combination and richness of these models determine the capability level of the DT-RAN. Hence, different capability levels are associated with different model compositions. For the basic capability level, the DT-RAN can simulate the static state and dynamic behavior

of physical networks, and accurately reflect the real network. For the advanced capability level, the DT-RAN can interact with the real network to collect real-time data from physical networks and use this feedback data for network configuration and control.

Based on the ITU-T recommendation Y.3090 [41], a reference framework of a digital twin network mainly consists of a data sharing module, a modeling module, and a digital twin management module. The data sharing module is responsible for collecting and storing various network data and providing data services and a unified interface to other modules. The modeling module completes data-based modeling and provides the obtained model for various network applications. The digital twin management module oversees the lifecycle management and visualization of the digital twin. Trustworthiness management for DT-RAN can be executed by the digital twin management module.

The trustworthiness of DT-RAN can be measured through trust evaluation [40]. Trust evaluation is a useful means to assessing the reliability of an object's behavior. When the management function (a trustor) needs to evaluate the trustworthiness level of the DT-RAN (a trustee), it can utilize the obtained monitoring metrics data associated with the monitoring objects as input to a trust calculation function or a trust inference model. The monitoring objects might include the data, model and network/computing resource related to the DT-RAN. The monitoring metrics differ depending on the monitored objects. For example, the model performance metrics can be the model accuracy assessed by comparing the simulation/predicted results and the ground truth results, while the network/computing resource monitoring metrics can include network bandwidth, data transmission rate, delay, and computing resource availability ratio.

In most cases, trust calculation functions use the weighted arithmetic operation or subjective logic model to calculate the trust value of a trustee based on the considered trust factors. The calculation or inference result can be seen as the trustworthiness level of the DT-RAN, which can serve as the basis of subsequent decision making.

For instance, consider a DT-RAN system that is evaluated based on three trust factors derived from monitoring metrics: model accuracy (weighted at 0.5), network bandwidth availability ratio (weighted at 0.3), and computing resource availability ratio (weighted at 0.2). Each factor is normalized to a scale of 0 to 10, with 10 being the highest.

Suppose the DT-RAN receives the following normalized ratings: model accuracy is 8.5, network bandwidth availability ratio is 7.0, and computing resource availability ratio is 9.0. The trustworthiness level of the DT-RAN can then be calculated using the weighted arithmetic operation:

TrustworthinessLevel = 0.5*8.5+0.3*7.0+0.2*9.0 = 4.25+2.1+1.8 = 8.15

This result indicates that the DT-RAN has a trustworthiness level of 8.15 out of 10.

High trustworthiness level means the DT-RAN output information/strategy/control can be reliably trusted and confidently applied to the real network. If the trustworthiness level of the DT-RAN drops below a certain threshold, management function might

need to downgrade DT-RAN capability level by reducing some models or replace/update/terminate certain models or revise data collection frequency or improve input data quality and so on.

Example Use Case:

Use Case 1: DT-RAN for Network Energy Saving

The DT-RAN plays a crucial role in achieving optimal network energy savings. It models diverse energy-saving strategies, simulating them within the DT-RAN to determine optimal energy saving strategy without compromising end-user QoS.

The Digital Twin virtual replica is mapped with real-network entities, environment, and their behavior. The real-time data is being collected from the network for predicting traffic patterns and current level of QoS to represent the network behavior and patterns. The DT-RAN would require the following inputs to replicate the real network in virtual models:

- Network topology and coverage KPI.
- Network & UE capability configurations (RATs, antennas, band support, etc.).
- User device density, traffic profiles.
- Specific service/user level SLAs.
- Traffic patterns/ KPIs.
- Power consumption datasets.

The DT-RAN represents the current state of the radio access network. The power saving requirement (e.g., power saving goal and user QoS targets) should be submitted to DT-RAN for checking the optimal state of the network for energy saving and analyzing its impact on network configuration changes, coverage, traffic migration and end-user QoS level. The DT-RAN will leverage AI/ML techniques and model representation (e.g., coverage prediction model, user re-distribution model, user experience prediction model, energy consumption prediction model) to arrive at an optimal state and can push the target parameters and configuration changes to the network to realize it. The DT-RAN output strategies for energy optimization in networks would include:

- Switching off a cell or carrier
- Switching off some of the RF ports
- Advanced sleep modes
- Efficient cloud resource management.

The adoption of the DT-RAN output energy saving strategy/control relies on the trustworthiness level of DT-RAN. The data, model and network/computing resource status all affect the trustworthiness of the DT-RAN. For example, the limited network transmission bandwidth will impact real-time data collection from the physical network, and the limited computing resource will impact the model construction and data

processing performance. All these will result in the corresponding decrease of the model performance. In order to ensure the desired DT-RAN trustworthiness level, these factors need to be monitored, including model performance, network/computing resource, etc.

For the model performance monitoring, the monitoring metrics can be the coverage prediction model accuracy, user experience prediction model accuracy and/or energy consumption prediction model accuracy. For the network resource status monitoring, the monitoring metrics can be the network element load and network transmission resource usage. Based on the monitored metrics, the DT-RAN trustworthiness can be evaluated by the weighted arithmetic operation or subjective logic model.

When the evaluated trustworthiness of DT-RAN is high, the energy-saving strategies/ controls output by DT-RAN can be applied on the real network; however, if the trustworthiness of DT-RAN is low, potential actions such as downgrading DTN capability level by reducing/terminating some models or replacing/updating certain complex models with simplified models of the DT-RAN can be taken into consideration to ensure the satisfactory performance of DT-RAN.

2.5 Intercommunication/Information Exchange

As stated earlier, a DT-RAN is an up-to-date virtual representation of the physical O-RAN system, asset, or process. It is a model continuously tuned to and reflecting the physical counterpart or twin while dynamically adapting to its operational and other changes. Furthermore, DT-RAN simulated model leverages data flow, AI/ML and analytics, and desired services to provide visualized insights and actionable decisions towards use cases such as new O-RAN scenario analysis and optimization with energy efficiency. This implies continuous and dynamic levels of interfacing and access, data flow and communication, as well as exposure and interactivity within the DT-RAN, and particularly across the DT-physical system.

A logical view of this system and levels of communication and interactivity, particularly across the physical and DT worlds, are shown in Figure 2-14. It is noteworthy that there can be a variety of proposed architectures based on applications, and the current state of research and development.

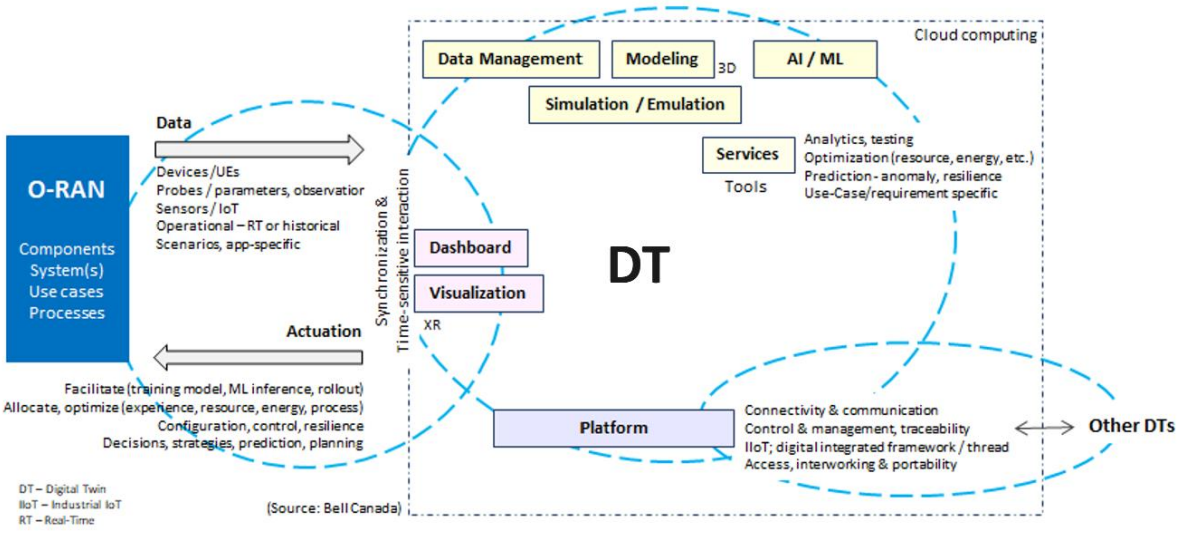


Figure 2-14: Communications within DT and across DT-Physical Systems

However, there is a common understanding that enabling efficient information exchange between various modules of DT-RAN as well as between DT-RAN and external entities, e.g., the underlying real network and associated network applications is crucial to the proliferation of scalable, agile, and highly reliable DT-RAN.

In order to train or calibrate (depending on the modeling technology being used) the models within the DT-RAN as well as to keep DT-RAN synchronized with the real network, data is to be collected from the real network. Capability of the existing O1 interface in O-RAN network can be enhanced to support the data collection requirement for DT-RAN. The enhancements can include – 1) more performance metrics (PM) to be exposed from the O-RAN network elements, 2) more instantaneous and event driven status reports, not only limiting to the statistical performance metrics, and 3) potentially improving bandwidth and latency on the O1 interface transport layer to support the demanding data collection requirements.

Once the DT-RAN is synchronized with the real network, other functions or applications can interact with the DT-RAN, e.g., training AI/ML models, doing 'what-if' analysis, predicting performance impact with a certain control command or policy update, and mitigating conflicts with other applications etc. Before a change is made to the real network, the change can be made in the DT-RAN first. Once the performance is validated and assured with DT-RAN, the change is forwarded to the real network to update the configuration or control of the real network.

3 Conclusion

DT Technology will play a pivotal role in the emerging 6G network – all the way from aiding in complex network planning to providing a risk-free environment with a highly-accurate and real-time executable experimental sandbox for trying various network configurations, executing what-if predictive analysis and innovating new network features like algorithms, protocols, topologies, etc. To realize the full potential of DTN, the right set of technologies and tools needs to be ensembled in creating different

modules of a DTN. In this research report, the authors have collaborated into a deep dive analysis of various building blocks and associated key enabling technologies for creating DT-RAN.

The three key pillars of a DT-RAN are identified as – Data, Modeling, and Interfaces. Data is the fuel for DT-RAN, and there are various types of data that need to be accumulated for creating faithful digital replica of complex RAN behavior. While data collected from the field (e.g., from network elements and physical environment) would be valuable, the sheer volume of data required for effectively training a DT model is likely not available purely from collected data. Other sources of data, for example synthetically generated data through various advanced AI/ML techniques or augmenting synthetic data with real-data would be critical towards meeting the quality, quantity and diversity needs for DT-RAN data. Managing data lifecycle is another important aspect of DT-RAN data. In particular, maintaining the information on DT-RAN data drift, i.e., the changes in the data distribution (either as input or output of the DT-RAN) compared to the real system data is crucial, and need to be collected and monitored on a regular basis. Based on the collected data drift information, timely updates of the DT-RAN with the new data collected from the real system would ensure the high fidelity and trustworthiness of DT-RAN over time.

Modeling DT-RAN involves various facets of the physical RAN, including network elements, surrounding physical environment, network subscribers and the end-to-end system behavior, and various technologies enabling each of these modeling aspects. Notable technology enablers for DT-RAN modeling include Analytical/Stochastic modeling, Simulator/Emulator based modeling, AI/ML based modeling, Ray Tracing (including differential ray tracing) and Graph Neural Networks. The degree of computational complexity varies across these different modeling approaches. But the common underlying principle enabling efficient DT-RAN modeling is the ability of fast computing, which can be aided by accelerated computing technologies. Key technology enablers for accelerated computing in DT-RAN manifest in various forms, e.g., acceleration hardware (e.g., GPU, DPU, TPU, NPU, FPGA, ASIC) and associated acceleration software (e.g., SW libraries accelerating mathematical computation), multi-GPU scaling and HPC, cloud computing and distributed/edge computing. Visualization of DT-RAN's simulation and data analytics through sophisticated tools like high end graphics (3D models, AR/VR), geospatial visualization, RF heatmaps, etc. would be crucial for deriving new and explainable insights into network's predictive behavior, what-if analysis, and configuration testing. While deriving valuable insights about real-network's behavior from its twin, it is critical to ensure the Digital Twin is in sync with the real-network and truly reflecting the network's current behavior – which makes performance monitoring of the DT-RAN an important criteria, enabled through trust evaluation.

Finally, the third pillar of DT-RAN, the interfaces connecting various components of DT-RAN as well as between the DT-RAN and its physical counterpart can be enabled by enhancing existing interfaces (like O1) in open RAN architecture and further research is needed to explore the detailed needs of these interfaces (e.g., type/volume of data to be exchanged, frequency of data transfer, targeted latency/bandwidth,

security aspects, etc.) and whether existing interfaces could serve the purpose through enhancements or new interfaces to be introduced.

The report also highlights a number of use cases to illustrate how these technology enablers can be combined together to facilitate DT-RAN deployment in various scenarios including DT-RAN for AI/ML training, testing and performance assurance, network planning, network energy saving, site-specific network optimization and network automation [42].

As a follow up of this research report, we will explore various functional, service, testing, security and architectural requirements of DT-RAN and some of the critical challenges in fulfilling these requirements, along with looking for potential solutions. Bringing all these research findings together will ultimately pave the way for understanding the nuances of building DT-RAN for emerging 6G networks, and can be used as a valuable guideline for extending DT-RAN for O-RAN centric networks.

References

[1] 3GPP TS 37.320, “Radio measurement collection for Minimization of Drive Tests (MDT); Overall description; Stage 2,” v18.1.0, Mar. 2024.

[2] 3GPP TR 22.804, “Study on Communication for Automation in Vertical domains,” v16.3.0, Nov. 2020.

[3] C. Zhou et al., “Digital Twin Network: Concepts and Reference Architecture,” IRTF Draft, Mar. 2024.

[4] A. Cabellos and C. Janz, “Performance-Oriented Digital Twins for Packet and Optical Networks,” IRTF Draft, Apr. 2024.

[5] X. Lin et. al., “6G Digital Twin Networks: From Theory to Practice,” IEEE Communications Magazine, vol. 61, no. 11, pp. 72-78, Nov. 2023.

[6] 3GPP TR 38.901, “Study on Channel Model for Frequencies from 0.5 to 100 GHz,” v17.0.0, Mar. 2022.

[7] RWS-230015, “Channel modelling to 7-24 GHz, sensing and RIS,” source: Nokia, Nokia Shanghai Bell.

[8] RWS-230173, “Channel modeling for new spectrum,” source: Qualcomm.

[9] RWS-230212, “Motivation and Proposal for FR3 Channel Model Study in Rel-19,” source: Samsung.

[10] RWS-230293, “Views on channel model in Rel-19,” source: ZTE, Sanechips.

[11] RWS-230421, “Study on realistic channel information collection for AI/ML in Rel-19,” source: CMCC.

[12] RWS-230120, “Study on Channel Modeling for Integrated Sensing and Communication,” source: MediaTek Inc.

[13] Z. Yun and M. F. Iskander, “Ray Tracing for Radio Propagation Modeling: Principles and Applications,” IEEE Access, vol. 3, pp. 1089-1100, 2015.

[14] J. Hoydis et. al., "Sionna RT: Differentiable Ray Tracing for Radio Propagation Modeling," arXiv:2303.11103, Mar. 2023.

[15] H. Choi et. al., “WiThRay: A Versatile Ray-Tracing Simulator for Smart Wireless Environments,” IEEE Access, vol. 11, pp. 56822-56845, 2023.

[16] 3GPP TR 22.804, “Study on Communication for Automation in Vertical domains (CAV),” v16.3.0, Nov. 2020.

[17] D. T. Ngo et. al., “Empowering Digital Twin for Future Networks with Graph Neural Networks: Overview, Enabling Technologies, Challenges, and Opportunities,” Future Internet, vol., 15, issue 12, pp. 377, 2023.

[18] P. Almasan et. al., “Network digital twin: Context, enabling technologies, and opportunities,” IEEE Communications Magazine, vol., 60, no. 11, pp. 22-27, Nov. 2022.

[19] M. Ferriol-Galmés et. al., “Building a digital twin for network optimization using graph neural networks,” Computer Networks, vol. 217, pp.109329, Nov. 2022.

[20] M. Besta et. al., “Neural graph databases,” in Proc. Learning on Graphs Conference (PMLR), 9-12 Dec. 2022.

[21] H. Zhang et. al., “Understanding GNN computational graph: A coordinated computation, IO, and memory perspective,” Proceedings of Machine Learning and Systems, vol. 4, pp. 467-484, 2022.

[22] Y. Yang et. al., “Graph Neural Network-Based Node Deployment for Throughput Enhancement,” IEEE Transactions on Neural Networks and Learning Systems, Early Access.

[23] J. Zhou et. al., “Graph neural networks: A review of methods and applications,” AI open, vol. 1, pp. 57-81, Ja. 2024.

[24] Y. Shen et. al., “Graph Neural Networks for Scalable Radio Resource Management: Architecture Design and Theoretical Analysis,” IEEE Journal on Selected Areas in Communications, vol. 39, no. 1, pp. 101-115, Jan. 2021.

[25] O. Orhan et. al., “Connection Management xAPP for O-RAN RIC: A Graph Neural Network and Reinforcement Learning Approach,” in Proc. 20th IEEE International Conference on Machine Learning and Applications (ICMLA), pp. 936-941, Pasadena, CA, USA, 2021.

[26] A. Bufort et. al., “Data-Driven Radio Propagation Modeling using Graph Neural Networks,” TechRxiv. November 07, 2023. [Online] Available: https://d197for5662m48.cloudfront.net/documents/publicationstatus/180305/preprint_pdf/69d6a92985df21734d1e5b4cfc1c4dbb.pdf

[27] A. Moayyedi et. al., “Generalizable GNN-based 5G RAN/MEC Slicing and Admission Control in Metropolitan Networks,” in Proc. IEEE/IFIP Network Operations and Management Symposium (NOMS), pp. 1-9, Miami, FL, USA, 2023.

[28] Y. Wu et. al., “Digital Twin Networks: A Survey,” IEEE Internet of Things Journal, vol. 8, no. 18, pp. 13789-13804, Sep. 15, 2021.

[29] A. Mozo et. al., “B5GEMINI: AI-Driven Network Digital Twin,” Sensors, vol. 22, issue 11, no. 4146, pp. 1-30, May 28, 2022.

[30] L. Hui et. al., “Digital Twin for Networking: A Data-Driven Performance Modeling Perspective,” IEEE Network, vol. 37, no. 3, pp. 202-209, May/Jun. 2023

[31] T. Liu et. al., “Digital-Twin-Assisted Task Offloading Based on Edge Collaboration in the Digital Twin Edge Network,” IEEE Internet of Things Journal, vol. 9, no. 2, pp. 1427-1444, Jan. 15, 2022.

[32] A. Taneja and S. Rani, “Energy Efficient Digital Twin Enabled Massive IoT Network with Use Case in Consumer Health,” IEEE Transactions on Consumer Electronics, Early access.

[33] F. He et. al., “An integrated mobile augmented reality digital twin monitoring system,” Computers, vol. 10, no. 8, pp. 99, Aug. 2021.

[34] L. Adreani et. al., “Design and develop of a smart city digital twin with 3d representation and user interface for what-if analysis,” in Proc. International Conference on Computational Science and Its Applications, pp. 531-548, Cham: Springer Nature Switzerland, Ju. 2023.

[35] A. Lee et. al., “A geospatial platform to manage large-scale individual mobility for an urban digital twin platform,” Remote Sensing, vol. 14, no. 3, pp. 723, 2022.

[36] G. Schroeder et al., “Visualising the digital twin using Web services and augmented reality,” in Proc. IEEE 14th Int. Conf. Ind. Informat. (INDIN), pp. 522–527, 2016.

[37] B. Erman and C. Martino, “Generative Network Performance Prediction with Network Digital Twin,” IEEE Network, vol., 37, no. 2, pp. 286-292, 2023.

[38] J. Zhao et. al., “Design and Application of a Network Planning System Based on Digital Twin Network,” IEEE Journal of Radio Frequency Identification, vol. 6, pp. 900-904, 2022.

[39] J. Nassif et. al., “Synthetic Data: Revolutionizing the Industrial Metaverse,” Springer Nature, 2024.

[40] J. Guo et al., “TFL-DT: A Trust Evaluation Scheme for Federated Learning in Digital Twin for Mobile Networks,” IEEE Journal on Selected Areas in Communications, vol. 41, no. 11, Nov. 2023.

[41] ITU-T recommendation Y.3090, “Digital twin network - Requirements and architecture.”

[42] O-RAN next Generation Research Group (nGRG) Contributed Research Report, “Digital Twin RAN Use Cases,” report id RR-2024-07, Jul. 31, 2024.

[43] X. Lin et. al., “A Primer on Generative AI for Telecom: From Theory to Practice,” arXiv:2408.09031, Aug. 16, 2024.

[44] T. Li et. al., “Differentiable Monte Carlo ray tracing through edge sampling,” ACM Transactions on Graphics (TOG), vol. 37, issue 6, article no. 222, pp. 1-11, Dec. 2018.

[45] B. Hughes et. al., “Generative Adversarial Learning for Machine Learning empowered Self Organizing 5G Networks,” in 2019 International Conference on Computing, Networking and Communications (ICNC), Honolulu, HI, USA: IEEE, Feb. 2019.

[46] M. N. A. Khatiman et. al., “Generation of Synthetic 5G Network Dataset Using Generative Adversarial Network (GAN),” in 2023 IEEE 16th Malaysia International Conference on Communication (MICC), Dec. 2023.