Conference Paper Title*

*Note: Sub-titles are not captured in Xplore and should not be used

1st Given Name Surname
dept. name of organization (of Aff.)
name of organization (of Aff.)
City, Country
email address or ORCID

2nd Given Name Surname
dept. name of organization (of Aff.)
name of organization (of Aff.)
City, Country
email address or ORCID

3rd Given Name Surname
dept. name of organization (of Aff.)
name of organization (of Aff.)
City, Country
email address or ORCID

4th Given Name Surname
dept. name of organization (of Aff.)
name of organization (of Aff.)
City, Country
email address or ORCID

5th Given Name Surname
dept. name of organization (of Aff.)
name of organization (of Aff.)
City, Country
email address or ORCID

6th Given Name Surname
dept. name of organization (of Aff.)
name of organization (of Aff.)
City, Country
email address or ORCID

***Abstract*—This document is a model and instructions for LaTeX. This and the IEEEtran.cls file define the components of your paper [title, text, heads, etc.]. *CRITICAL: Do Not Use Symbols, Special Characters, Footnotes, or Math in Paper Title or Abstract.**

***Index Terms*—component, formatting, style, styling, insert**

I. Introduction

This document is a model and instructions for LaTeX. Please observe the conference page limits.

II. Ease of Use

A. Maintaining the Integrity of the Specifications

The IEEEtran class file is used to format your paper and style the text. All margins, column widths, line spaces, and text fonts are prescribed; please do not alter them. You may note peculiarities. For example, the head margin measures proportionately more than is customary. This measurement and others are deliberate, using specifications that anticipate your paper as one part of the entire proceedings, and not as an independent document. Please do not revise any of the current designations.

III. AI Toolkits and Potential Applications for Mobile Networks

The integration of Artificial Intelligence (AI) into the Radio Access Network (RAN) represents a paradigm shift from traditional, human-driven optimization to intelligent, autonomous network control. However, bridging the gap between advanced AI models and the real-time, high-stakes environment of a mobile network requires a new class of foundational software. AI toolkits for the RAN are specialized frameworks designed to address this challenge, providing the essential infrastructure to develop, deploy, and manage AI-driven applications within the complex cellular architecture.

The primary goal of these toolkits is to streamline the entire AI workflow by addressing three critical pillars:

1. GPU-Acceleration: The RAN operates under extremely strict latency constraints. These toolkits provide optimized libraries that leverage the parallel processing power of Graphics Processing Units (GPUs) to ensure that sophisticated AI algorithms can execute within the real-time budget of the RAN.
2. Data Processing and Management: A modern RAN is a prolific source of data. AI toolkits offer robust data pipelines to ingest, aggregate, and pre-process this high-velocity data, transforming it into a clean, model-ready format.
3. AI Lifecycle Management (MLOps): The operational challenge lies in managing thousands of models distributed across a live network. These toolkits incorporate MLOps principles tailored for telecommunications, automating the end-to-end lifecycle of AI models to ensure they remain robust, secure, and effective over time.

In essence, AI toolkits serve as the critical enabling layer, abstracting away the underlying hardware and data complexity to empower developers and operators to build and operate a truly intelligent RAN.

A. AI toolkits

NVIDIA provides several advanced AI toolkits that accelerate the design, testing, and operation of current and future wireless communication systems, including 5G and 6G.

1) RAN-Specific Platforms & SDKs: While the concept of an AI toolkit is broad, several industry players offer specific platforms and SDKs. An exemplary platform is the NVIDIA Aerial SDK, a complete software stack for building high-performance, GPU-accelerated, cloud-native 5G and 6G virtualized RANs (vRANs). Its architecture is centered on two core libraries:

Identify applicable funding agency here. If none, delete this.

- cuBB (CUDA Baseband): Regarded as the “heart of Aerial,” it acts as the signal processing engine by implementing the entire L1-L2 stack (cuPHY and cuMAC) on the GPU, ensuring high throughput by keeping all processing within high-performance GPU memory.
- cuVNF (CUDA Virtual Network Functions): This SDK serves as the I/O engine, providing optimised input/output directly to GPU memory from GPUDirect-capable Network Interface Cards (NICs).

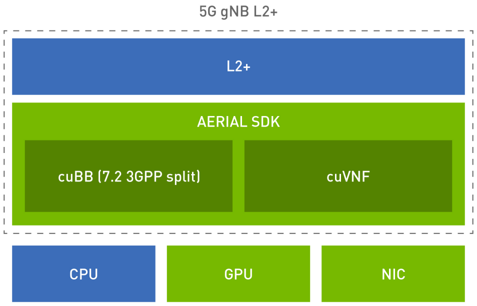

Fig. 1. NVIDIA Aerial SDK platform for GPU-accelerated, cloud-native vRANs.

The broader market of established vendors is also integrating AI. Qualcomm focuses on powerful hardware like the X100 Accelerator Card for inline acceleration. Ericsson embeds AI deeply into its solutions via its Intelligent RAN Automation portfolio. Nokia offers both a commercial RAN Intelligent Controller (RIC) and integrates AI into its AirScale baseband portfolio.

2) The O-RAN Architectural Framework: The O-RAN Alliance defines a disaggregated and intelligent architecture for the RAN, centered on the RAN Intelligent Controller (RIC). The RIC functions as the network’s ”operating system,” enabling AI-driven control via a modular, app-based model. It is split into the Non-Real-Time RIC (for network-wide policies via rApps) and the Near-Real-Time RIC (for low-latency control via xApps).

3) Simulation and Data Generation Toolkits: The development of robust AI models is critically dependent on large, high-quality datasets. Simulation toolkits fill this gap.

- NVIDIA Sionna: An open-source, Python-based library for link-level simulation. It is designed for communications research and enables the rapid prototyping of AI/ML systems for the physical layer. It includes modules for ray tracing (Sionna RT), link-level simulation (Sionna PHY), and system-level simulation (Sionna SYS). A key component is the Sionna Research Kit (SRK), which enables the deployment of trained models into a real software-defined 5G network on the NVIDIA Jetson AGX Orin platform.
- NVIDIA Aerial Omniverse Digital Twin (AODT): A next-generation, system-level simulation platform for R&D on 5G/6G systems. AODT applies ray-traced channels to the NVIDIA Aerial RAN platform, simulating the system-level performance of an actual network deployment without abstractions. It is a unique tool to benchmark performance and explore ML algorithms under real-world conditions.

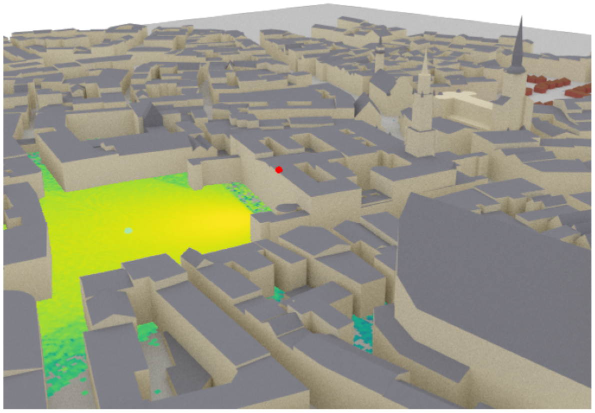

Fig. 2. NVIDIA Sionna simulation toolkit for AI-native air interface research.

4) General Purpose & Emerging Toolkits: A new frontier in network automation is the application of Large Language Models (LLMs) and Generative AI (GenAI). Toolkits involving models like Gemini, coupled with the Retrieval-Augmented Generation (RAG) technique, are being explored for complex network operations like automated root cause analysis, configuration script generation, and interactive troubleshooting.

B. Technical feasibility

The introduction of the AI toolkits described raises critical questions about their practical implementation and technological readiness. This section assesses the technical feasibility, focusing on their performance, data dependency, maturity, and operational robustness.

1) Model Performance and Algorithmic Efficiency: The feasibility of toolkits like the NVIDIA Aerial SDK and the O-RAN RIC hinges on their ability to execute AI algorithms within the RAN’s stringent latency budgets. For platforms such as Aerial, which move L1 processing to GPUs, the key feasibility question is whether this software-defined approach can consistently match or exceed the performance of traditional hardware while simultaneously running AI models for real-time tasks like beam management. Similarly, for the O-RAN framework, the performance of an xApp is constrained by the processing capacity of the Near-RT RIC platform it runs on and the latency of the E2 interface, making algorithmic efficiency a non-negotiable requirement for any viable application.

2) Data Availability and Accessibility: The AI toolkits discussed are fundamentally dependent on data. For toolkits aimed at the physical layer, like NVIDIA Sionna, feasibility is enhanced by its ability to synthetically generate high-fidelity training data. For emerging GenAI toolkits, feasibility depends on creating robust RAG pipelines that can interpret diverse operator data sources.

3) Technology Maturity and Integration: While individual technologies are maturing, their integration poses a significant technical challenge. The NVIDIA Aerial SDK is a mature product, but its feasibility depends on seamless integration into a multi-vendor vRAN ecosystem. The O-RAN framework itself represents a major integration challenge focused on interoperability.

However, the feasibility of achieving large-scale integration is strongly validated by successes in other critical infrastructure sectors. The creation of national-scale digital twins for railway networks—as implemented by Deutsche Bahn and proposed for Malaysia's economic development—demonstrates the maturity of platforms like NVIDIA Omniverse. These projects prove it is technically feasible to fuse massive, heterogeneous data streams, from real-time IoT sensor data to geographical information systems (GIS) like OpenStreetMap, into a cohesive, synchronized virtual model. This precedent supports the argument that achieving a similar level of integration for a national mobile network is technically feasible.

4) Security, Reliability, and Scalability: For these toolkits to be technically feasible, they must meet telco-grade requirements. The reliability of a software-defined stack like NVIDIA Aerial must be proven to match the "five-nines" (99.999%) availability of traditional hardware. The scalability of managing thousands of unique xApp instances across a national network poses a significant MLOps challenge.

C. Potential applications in mobile networks

The surveyed toolkits facilitate a range of high-impact use cases for building intelligent mobile networks:

- AI-Native Air Interface: The NVIDIA Aerial SDK enables the development of highly programmable and scalable 5G vRANs where AI/ML frameworks can be seamlessly integrated for real-time signal processing.
- System-Level R&D and Optimization: The NVIDIA AODT allows researchers to benchmark system performance, explore ML-based algorithms, and optimize network planning using physically accurate simulations without abstractions.
- **Network-Scale Digital Twins for Operations:** Platforms like **NVIDIA Omniverse** are enabling digital twins of national infrastructure. For instance, Deutsche Bahn has already modeled its railway network, while **proposals for Malaysia's railway system highlight the strong economic drivers** for such projects, aiming to boost GDP and operational efficiency. This large-scale simulation concept is directly applicable to mobile networks for optimizing national 5G/6G cell placement, managing resources, and planning capacity by integrating network data with geographical information from sources like OpenStreetMap.
- Rapid Prototyping and Deployment: NVIDIA Sionna accelerates research into 6G by enabling fast, GPU-accelerated modeling of communication systems, while its Sionna Research Kit (SRK) bridges the gap from research to reality by enabling deployment into a real software-defined 5G network.

D. Deployment constraints in mobile networks

The practical deployment of the surveyed AI toolkits is constrained by a series of specific hardware and software requirements that must be met for successful implementation and operation. These are not theoretical challenges but concrete prerequisites essential for the toolkits to function correctly within a mobile network. Successful deployment hinges on the availability of powerful and specific hardware, including particular NVIDIA GPUs, qualified server systems like the Dell R750, and substantial system memory. Furthermore, the software environment must be precisely configured, mandating specific versions of operating systems, drivers, and libraries like CUDA and TensorFlow, often necessitating the use of containerized environments.

Hardware: Requires specific NVIDIA GPUs with substantial vRAM (e.g., RTX 6000 Ada with 12GB+ for frontend; A100/H100 with 48GB+ for backend). Qualified systems include Dell R750 servers with Intel Xeon Gold CPUs and 512GB DDR4 Memory.

Software: The backend requires Ubuntu 22.04. The frontend supports Ubuntu 22.04 and Windows 11. Specific NVIDIA driver versions are required (e.g., 560.35.05 for Linux backend).

Hardware: Runs on NVIDIA-certified EGX servers. Qualified platforms use GPUs like the NVIDIA A100x, DPUs like the Mellanox CX6-DX, and CPUs such as the Intel Xeon Gold 6336Y. Memory ranges from 96GB to 512GB of DDR4 RAM.

Software: Requires a specific stack, including Ubuntu 20.04 with a low-latency kernel, CUDA 11.7 Toolkit, and Kubernetes version 1.23.

Hardware: While GPU-accelerated for performance, Sionna can also run on a CPU. The Sionna Research Kit (SRK) component is specifically designed for the NVIDIA Jetson AGX Orin platform.

Software: Requires Python 3.8-3.12 and TensorFlow 2.14-2.19. Ubuntu 24.04 is the recommended operating system. A Docker container or Python virtual environment is highly recommended.

BibTeX does not work by magic. It doesn't get the bibliographic data from thin air but from .bib files. If you use BibTeX to produce a bibliography you must send the .bib files.

LaTeX can't read your mind. If you assign the same label to a subsubsection and a table, you might find that Table I has been cross referenced as Table IV-B3.

LaTeX does not have precognitive abilities. If you put a `\label` command before the command that updates the counter it's supposed to be using, the label will pick up the last counter to be cross referenced instead. In particular, a `\label` command should not go before the caption of a figure or a table.

Do not use `\nonumber` inside the `{array}` environment. It will not stop equation numbers inside `{array}` (there won't be any anyway) and it might stop a wanted equation number in the surrounding equation.

E. Some Common Mistakes

- The word "data" is plural, not singular.
- The subscript for the permeability of vacuum μ_0, and other common scientific constants, is zero with subscript formatting, not a lowercase letter "o".
- In American English, commas, semicolons, periods, question and exclamation marks are located within quotation marks only when a complete thought or name is cited, such as a title or full quotation. When quotation marks are used, instead of a bold or italic typeface, to highlight a word or phrase, punctuation should appear outside of the quotation marks. A parenthetical phrase or statement at the end of a sentence is punctuated outside of the closing parenthesis (like this). (A parenthetical sentence is punctuated within the parentheses.)
- A graph within a graph is an "inset", not an "insert". The word alternatively is preferred to the word "alternately" (unless you really mean something that alternates).
- Do not use the word "essentially" to mean "approximately" or "effectively".
- In your paper title, if the words "that uses" can accurately replace the word "using", capitalize the "u"; if not, keep using lower-cased.
- Be aware of the different meanings of the homophones "affect" and "effect", "complement" and "compliment", "discreet" and "discrete", "principal" and "principle".
- Do not confuse "imply" and "infer".
- The prefix "non" is not a word; it should be joined to the word it modifies, usually without a hyphen.
- There is no period after the "et" in the Latin abbreviation "et al.".
- The abbreviation "i.e." means "that is", and the abbreviation "e.g." means "for example".

F. Authors and Affiliations

The class file is designed for, but not limited to, six authors. A minimum of one author is required for all conference articles. Author names should be listed starting from left to right and then moving down to the next line. This is the author sequence that will be used in future citations and by indexing services. Names should not be listed in columns nor group by affiliation. Please keep your affiliations as succinct as possible (for example, do not differentiate among departments of the same organization).

G. Identify the Headings

Headings, or heads, are organizational devices that guide the reader through your paper. There are two types: component heads and text heads.

Component heads identify the different components of your paper and are not topically subordinate to each other. Examples include Acknowledgments and References and, for these, the correct style to use is "Heading 5". Use "figure caption" for your Figure captions, and "table head" for your table title. Run-in heads, such as "Abstract", will require you to apply a style (in this case, italic) in addition to the style provided by the drop down menu to differentiate the head from the text.

Text heads organize the topics on a relational, hierarchical basis. For example, the paper title is the primary text head because all subsequent material relates and elaborates on this one topic. If there are two or more sub-topics, the next level head (uppercase Roman numerals) should be used and, conversely, if there are not at least two sub-topics, then no subheads should be introduced.

H. Figures and Tables

a) Positioning Figures and Tables: Place figures and tables at the top and bottom of columns. Avoid placing them in the middle of columns. Large figures and tables may span across both columns. Figure captions should be below the figures; table heads should appear above the tables. Insert figures and tables after they are cited in the text. Use the abbreviation "Fig. 1" or "Fig. 2", even at the beginning of a sentence.

TABLE I
TABLE TYPE STYLES

Table Head	**Table Column Head**		
	Table column subhead	***Subhead***	***Subhead***
copy	More table copy[a]		

[a] Sample of a Table footnote.

IV. Conclusion

The integration of artificial intelligence is a critical component of the roadmap to 6G wireless networks [1]. This has

Fig. 3. Sample figure referenced in the text.

spurred the development of advanced toolkits from major industry players. For example, established vendors like Ericsson are offering platforms for intelligent RAN automation [2].

At a more granular level, NVIDIA provides a suite of tools for building GPU-accelerated RAN. This includes detailed software components for the baseband and physical layers, such as the Aerial cuBB and cuPHY [3], [4]. Furthermore, to facilitate the research and training of AI models for these new systems, simulation libraries like NVIDIA Sionna have become essential [5].

REFERENCES

[1] K. B. Letaief, W. Chen, Y. Shi, J. Zhang, and Y.-J. A. Zhang, "The roadmap to 6g: Ai empowered wireless networks," *IEEE Communications Magazine*, vol. 57, no. 8, pp. 84–90, Aug. 2019.

[2] Ericsson. (2025) Intelligent ran automation. Accessed: Jul. 18, 2025. [Online]. Available: https://www.ericsson.com/en/ran/intelligent-ran-automation

[3] NVIDIA. (2025) Aerial cubb quickstart overview. Accessed: Jul. 18, 2025. [Online]. Available: https://docs.nvidia.com/aerial/archive/cuda-accelerated-ran/24-1/aerial_cubb/cubb_quickstart/cubb_quickstart_overview.html

[4] ——. (2025) Aerial sdk cuphy overview. Accessed: Jul. 18, 2025. [Online]. Available: https://docs.nvidia.com/aerial/aerial-cuphy/current/text/overview.html

[5] ——. (2025) Sionna: Link-level simulation library for ai-native air interface research. Accessed: Jul. 18, 2025. [Online]. Available: https://nvlabs.github.io/sionna/