AI-RAN: Artificial Intelligence – Radio Access Networks

Frequently Asked Questions

March 2025

Document History

Version	Date	Authors	Description of Change
01	March 18, 2025	Shuvo Chowdhury, Rajesh Gadiyar, Emeka Obiodu	Initial release

Table of Contents

List of Figures

1. Overview

1.1 What is AI-RAN?

AI-RAN (Artificial Intelligence - Radio Access network) is a technology that enables full integration of AI into radio access network hardware and software to enable new AI services and monetization opportunities, in addition to the transformative gains in network utilization, spectral efficiency and performance.

The underlying infrastructure for AI-RAN is built using a completely homogeneous general purpose, accelerated computing platform, without any RAN specific hardware components, so that it can run both cellular and AI workloads concurrently with deterministic performance for each. It embodies cloud-native principles such as on-demand scaling, multi-tenancy, and containerization of both workloads.

The software for AI-RAN is built using fully software defined and AI-native principles to allow containerization and acceleration of AI and RAN workloads, ensuring full benefits of underlying accelerated computing infrastructure.

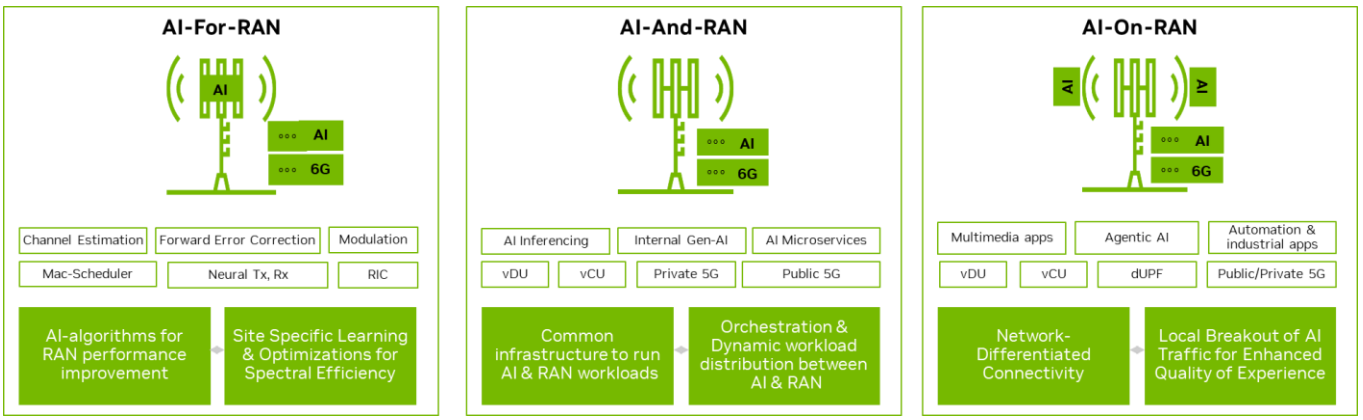
With this accelerated and unified hardware-software foundation, AI-RAN enables the deployment of 5G/6G RAN and AI workloads on a shared, distributed, and accelerated cloud infrastructure. It converts the RAN infrastructure from a single-purpose to multi-purpose cloud infrastructure.

There are three specific areas of AI integration into the RAN, as outlined by the AI-RAN Alliance – a community of telecom companies and academia with the mission to drive innovation and adoption of AI-RAN.

- AI and RAN (also referred to as AI with RAN): using a common shared infrastructure to run both AI and workloads, with the goal to maximize utilization, lower Total Cost of Ownership (TCO) and generate new AI-driven revenue opportunities.
- AI for RAN: advancing RAN capabilities through embedding AI/ML models, algorithms and neural networks into the radio signal processing layer to improve spectral efficiency, radio coverage, capacity and performance.
- AI on RAN: enabling AI services on RAN at the network edge to increase operational efficiency and offer new services to mobile users. This turns the RAN from a cost centre to a revenue source.

AI-RAN furthers the goals of Open-RAN, by leveraging a fully software-defined general purpose platform architecture, that enables open interfaces, to deliver flexibility, interoperability and cost-efficiency for the RAN.

Figure 1. The domains of AI-RAN



1.2 Why AI-RAN?

AI-RAN lays the technology foundation for the telecommunications industry to integrate the rapid advancements in AI technologies into the cellular telecommunications roadmap.

The surge in AI and generative AI applications is creating increased demands on cellular networks, driving demand for AI inferencing at the edge and necessitating new approaches to handle these workloads.

At the same time, advances in AI-based radio signal processing techniques are showing compelling results versus traditional techniques, and promising transformative gains in radio efficiency and performance.

As the industry begins its 6G journey, AI-RAN built with general purpose Commercial-Off-The-Shelf (COTS) servers and software defined acceleration, provides enhanced capabilities to process increased AI and non-AI traffic efficiently, compared to traditional RAN systems that are based on purpose-built hardware, whether it be custom Application-Specific Integrated Circuit (ASICs) or System on Chips (SoCs) with embedded accelerators.

AI-RAN creates new revenue opportunities from hosting AI workloads and enables AI to be integrated into the operations of the RAN to optimize network performance, automate management tasks, and enhance overall user experience.

1.3 What are the benefits of AI-RAN?

AI-RAN enables the deployment of 5G RAN and AI workloads on a shared, distributed, and accelerated cloud infrastructure thereby addressing the two key challenges Communication Service Providers (CoSPs) have had for a long time:

1. Average infrastructure utilization is low leading to lower return on investment (ROI).
2. Monetization of RAN-only services has limited upside as it is seen as a basic accessible service, yet the traffic is increasing and acquiring new spectrum or cell sites to serve growing traffic is expensive.

AI-RAN's core mission is maximizing the ROI for service providers by delivering the following key benefits to CoSPs:

- Maximizing utilization of their infrastructure resulting in lower TCO.
- New monetization opportunities via hosted AI services resulting in increased revenues.
- Improving spectral efficiency, energy efficiency and performance using AI techniques embedded into radio signal processing.
- Future proofing their infrastructure investments.

1.4 Why is AI-RAN transformational for Communications Service Providers (CoSP)?

AI-RAN is transformational to CoSPs because it:

- Delivers highest cell density, throughput and spectral efficiency for RAN, while ensuring carrier-grade and deterministic performance of RAN workloads.
- Enables CoSPs to dynamically assign unused RAN capacity for AI workloads, increasing the overall ROI through new monetization opportunities.
- Enhances the energy efficiency of fully loaded system.
- Future proofs CoSP's infrastructure investments with the ability to deploy ongoing improvements (RAN and AI) via new software releases, using Continuous Integration/Continuous Delivery (CI/CD) approach on the shared accelerated hardware platform, including a future software upgrade to 6G.

1.5 What are the building blocks for AI-RAN?

The key building blocks for AI-RAN include the following:

- Multi-purpose cloud native infrastructure - supports any RAN, any Cloud-Native Network Function (CNF), any Business Support Systems / Operations Support Systems (BSS/OSS) based internal AI workloads or any external AI workloads.
- Software defined architecture using COTS servers - no fixed function or purpose-built hardware.
- General purpose acceleration - that can accelerate multiple workloads.
- Multi-tenant and multi-workload capable design; both AI and RAN as first-class citizens, each with deterministic performance as per requirements.
- Scalable and fungible infrastructure; same servers can be used for any workload optimally with software reconfiguration and same homogenous infrastructure can be used for any deployment scenario including Centralised RAN (C-RAN), Distributed RAN (D-RAN) and Massive Multiple Input Multiple Output (mMIMO) variants, not requiring bespoke infrastructure for each use-case.

2. Reference Architecture

2.1 Is there a well-defined reference architecture for deploying AI-RAN?

AI-RAN is a fully software-defined, general-purpose solution where both AI and RAN workloads are treated as first-class citizens. Any hardware platform designed for this system should be capable of accelerating and supporting each workload independently, while enabling AI-and-RAN, AI-for-RAN and AI-on-RAN capabilities.

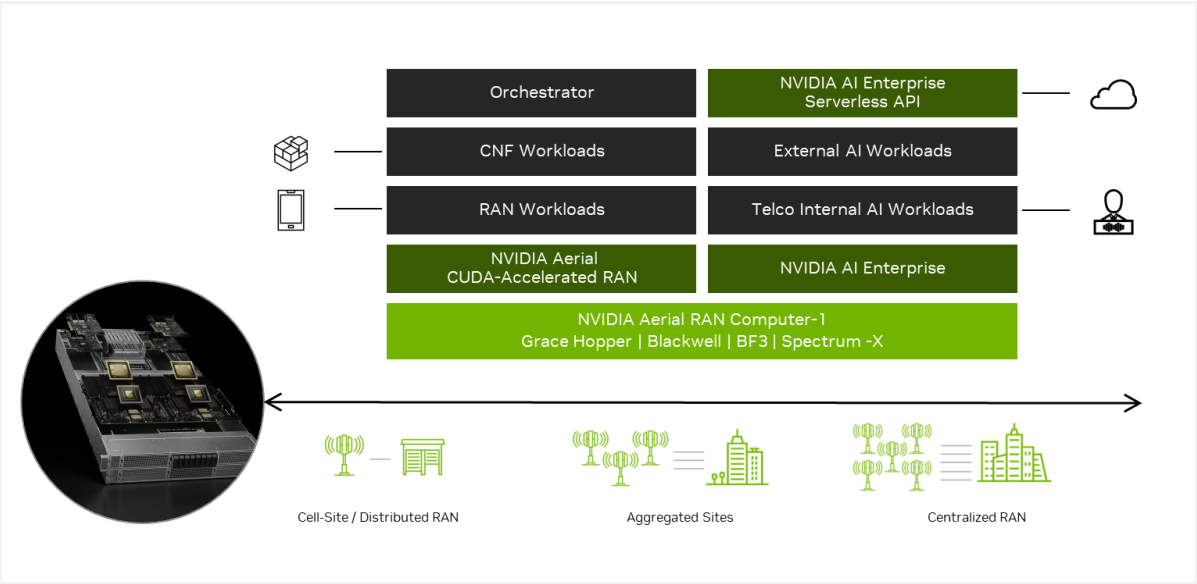
NVIDIA has worked with our partners to define, build and validate NVIDIA Cloud Partners (NCP) Telco Reference Architecture (RA). The goal of this RA is to create a blueprint that can drive rapid deployment of AI-RAN for the CoSP customers. The key ingredients of this RA include:

- Standard rack mounted Telco servers.
- NVIDIA MGX GH200 based Original Equipment Manufacturer (OEM) server platforms.
- Spectrum X compliant Fronthaul aggregation switches and Network Interface Controllers (NICs) - Spectrum Switches and Bluefield 3 (BF3) Data Processing Units (DPUs) - supporting timing requirements for RAN fronthaul and optimized AI ethernet capabilities.

The detailed design of this Reference Architecture (RA) (Figure 2) is explained in the ensuing section. NVIDIA has built and validated this RA. Additionally, some of our partners have successfully utilized the NCP Telco RA for AI-RAN field trials.

A key component to enable external AI workloads to be processed on-demand via the AI-RAN infrastructure is made possible by the NVIDIA AI Enterprise Serverless Application Programming Interface (API) that can fetch workloads from other data centers.

Figure 2. NVIDIA AI-RAN Reference Architecture



2.2 What is NVIDIA's AI-RAN Reference Architecture (RA) and how does it future proof infrastructure investments?

AI-RAN built with NVIDIA MGX GH200 servers and NVIDIA BF3, CX7/CX8 NICs and Spectrum-X switch fabric is fully programable and scalable. It can accommodate the evolving landscape of AI applications and the evolution to future 6G Networks with a software upgrade on the same hardware.

The AI-RAN reference architecture is built on the foundational principles of high-performance, scalability, and modularity in AI and RAN convergence.

To guide AI-RAN deployments, a solution blueprint consisting of a standard datacenter rack with AI-RAN servers comprising CPUs, GPUs, DPUs, solid-state drives (SSDs) and ethernet switch based networking fabric is shown in Figure 3 [1].

The schematic provides a reference architecture for CoSPs to deploy the next-generation, software defined and accelerated data center for AI-RAN, addressing the computational needs of AI and RAN workloads together.

Cloud-native accelerated compute is at the core of this reference architecture, enabling rapid deployment of AI-RAN systems with varying degrees of scaling and computing demands as per the RAN traffic and AI workload coming at the telco distributed data centers (such as central offices and mobile switching offices) over time.

This end-to-end AI-RAN deployment blueprint includes key components such as radio units (RUs), fronthaul (FH) network, distributed units (DUs) and optionally, Centralized Units (CUs) and Core Network (CN), all running on AI-RAN servers.

Note that, for simplicity, Figure 3 depicts fronthaul network topology connecting RUs to a single AI-RAN server (i.e., many-to-one mapping), whereas in practical deployments the connections between RUs and AI-RAN servers will be many-to-many.

For seamless flow of AI and RAN traffic through the same infrastructure, the networking fabric is divided into two parts, viz., compute fabric (between RUs and AI-RAN servers) and converged fabric (between AI-RAN servers and internet).

The compute fabric distributes the RAN workload via fronthaul across AI-RAN servers i.e. east-west (E-W) traffic flow. The converged fabric carries traffic for the combined RAN and AI workloads to and from the AI-RAN servers via midhaul/backhaul (north-south (N-S) traffic flow), and also provides connectivity to the wired networks for the AI traffic that is not originating from the wireless network.

Figure 3. Schematic of AI-RAN reference architecture [1]

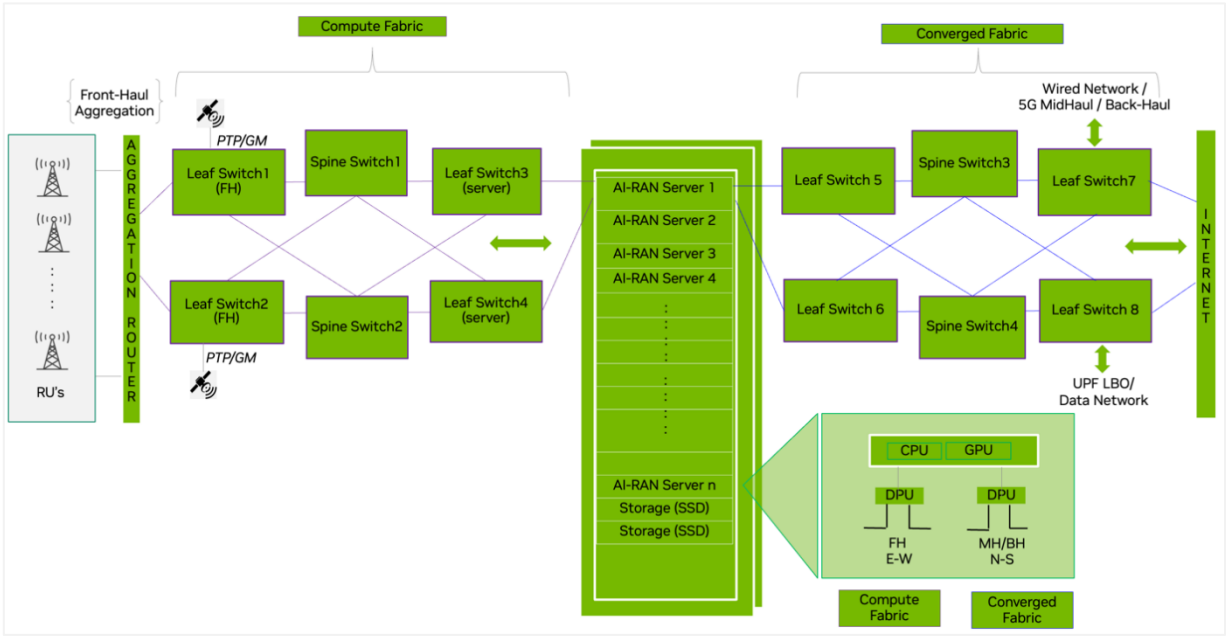


Figure 3 shows both these fabrics with a minimal two-spine, four-leaf tree topology, which can further scale in real-world deployments. Within the compute fabric, fronthaul connections coming from the RUs are aggregated in a cell site/transport aggregation router and connected to the AI-RAN servers via a 'spine-leaf' networking fabric. This two-switching layer architecture is commonly used in datacenter networking topology for scalability, redundancy, performance, and simplified network management.

In a typical spine-leaf mesh, leaf switches connect directly to network edge endpoints (e.g., servers and other edge devices), aggregating traffic from them before sending to the spine layer, while spine switches form the core of the networking fabric, routing traffic between leaf switches. The compute fabric comprises two types of leaf switches, viz., fronthaul leaf pair switches serving as access points for RUs (leaf switches 1-2 and

server leaf pair switches 3-4) connecting spine layer to edge AI-RAN servers. Each fronthaul leaf switch distributes timing via precision time protocol grandmaster (PTP/GM), the primary source of timing synchronization within the compute fabric of the network. Utilizing PTP protocol, the fronthaul leaf switches distribute precision timing information to RUs connected to the fronthaul network as well as to the DUs in the AI-RAN servers via lower layer split configuration 3 (LLS-C3) synchronization topology as per O-RAN fronthaul specification. The mesh topology created by fronthaul/server leaf switches interconnecting via spine layer creates a highly scalable and redundant network architecture in compute fabric.

While each AI-RAN server connects to the compute fabric at the front end (i.e., towards the fronthaul network), its backend connects to converged network leaf pairs (i.e., leaf switches 5-8) interconnected via a mesh of spine switches (i.e., spine switches 3-4). The converged fabric connects the AI-RAN servers to midhaul, backhaul, or internet depending on whether the AI-RAN servers are hosting only DU, combined DU and CU, or combined DU, CU and CN. For example, AI-RAN servers hosting only DU may connect via midhaul towards CU, whereas AI-RAN servers cohosting DU and CU could connect via backhaul towards user plane function-local breakout (UPF LBO) or towards UPF in CN. A centralized DU+CU+CN running on AI-RAN servers, on the other hand, would connect to the internet (via N6 interface) through the converged fabric, as illustrated in Figure 3.

Next, zooming into the AI-RAN servers in Figure 3, we explore the software stack built upon these servers to support the AI and RAN multi-tenancy on the same platform. Figure 2 illustrates various components of the software stack. It is designed to be cloud-native, with a commercial grade cloud operating system (e.g., Kubernetes) offering dynamic resource orchestration and infrastructure management. Cloud operating system hosts computing platforms and application programming interface (API) models like compute unified device architecture (CUDA) as well as networking platforms and API models like datacenter infrastructure on-a-chip architecture (DOCA) to efficiently run various RAN and AI applications aided by accelerated compute.

For the RAN stack, the DU, CU and CN are orchestrated by service management and orchestration (SMO) entity supporting multitudes of cells and RAN applications, whereas for the AI stack, various software components work in concert under API cluster agent that monitors and manages AI server workload in Kubernetes clusters. For the AI stack, the fundamental building blocks comprise AI application software frameworks and AI inferencing microservices (e.g., NVIDIA inferencing microservice (NIM), and NeMO framework), and industry-standard APIs to connect these components with various AI applications (e.g., text, speech, video, image) running on this platform either natively or through serverless APIs. An overarching end-to-end (E2E) orchestrator simultaneously works with RAN SMO and API cluster agent to track resource utilization and orchestrate RAN workload and AI inferencing requests on the same shared hardware, enabling multi-tenancy while maintaining desired quality-of-service and quality-of-experience requirements for RAN.

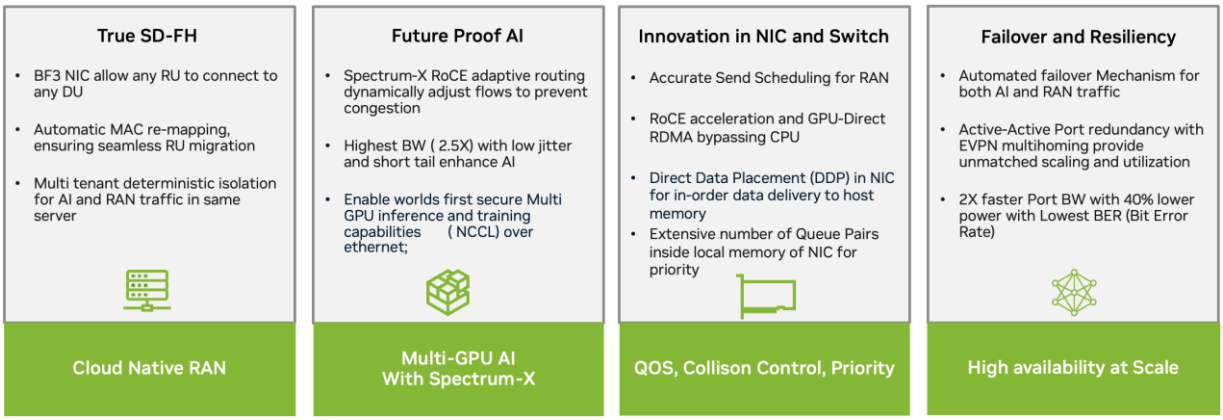
With the AI-RAN reference architecture and the associated software stack, a complete AI-RAN deployment blueprint is available for network operators to enable AI with RAN in the same infrastructure addressing various use case scenarios.

3. Switches and NIC Cards

3.1 Why does AI-RAN use specific types of switches and NIC cards?

AI-RAN when deployed in a multi-tenant datacenter to host C-RAN and AI services together, utilizes NVIDIA Spectrum-X technology to enable high performance networking fabric.

Figure 4. AI-RAN Fabric – Software Defined, High Performance, Programmable, and Scalable



The Spectrum-X Networking Platform, featuring Spectrum-4 switches and BF3 SuperNIC, is the world's first Ethernet fabric built for AI, accelerating generative AI network performance by 1.6X over traditional Ethernet fabrics. It is optimized for AI computing which has different networking than Cloud computing. Figure 5 highlights the difference between ethernet based cloud vs AI computing ethernet networking need.

Figure 5. Characteristics of traditional Cloud Networks and Networks for AI

Traditional Ethernet-based Clouds	AI Computing Ethernet Networking
Loosely coupled applications	Distributed tightly coupled processing
Low bandwidth TCP flows and utilization	High bandwidth RoCE flows and utilization
High jitter tolerance	Low jitter tolerance
Heterogeneous traffic, statistical multi-pathing	Bursty network capacity, elephant flows

Additionally, Spectrum Switch along with NVIDIA BF3 implements the following functions for AI-RAN:

1. Fronthaul Network with Precision Time Protocol (PTP) to connect RU to cloud hosted DU.
2. Software defined fronthaul capabilities to elastically orchestrate RAN and AI capabilities on the same infrastructure.
3. Medium Access Control (MAC) address remapping to ensure seamless DU migration.
4. Ethernet Virtual Private Network (EVPN) multi-homing and Active-Active redundancy for AI and RAN traffic with the help of Accelerated Linux bridge, enhanced NIC FW supporting PTP redundancy.
5. Security offloads and Service Function Chaining.
6. Mid-haul Network (towards CU) and Backhaul Network (towards 5GC).
7. AI Ethernet (East-West network traffic within the datacenter and North-South network traffic from/to the internet) – with features for Low Latency, Data Plane Acceleration, Congestion Control etc. for both RAN and AI applications.
8. Efficient Storage Networking - 400Gb/s Remote Direct Memory Access (RDMA) over Converged Ethernet (RoCE), RoCE adaptive routing and packet reordering.

AI-RAN spans both C-RAN and D-RAN deployments.

AI-RAN when deployed as a D-RAN may not utilize all the performance capabilities of Spectrum-X as described above. A typical D-RAN deployment consists of a single server (or a small number of servers) at a cell-site and is typically a power and thermally constrained environment.

D-RAN networking requires PTP and accurate scheduling of FH traffic. Additionally, when the platform is underutilized for RAN, it can be used for running some AI, typically, AI inference functions.

For D-RAN, NVIDIA offers two networking solutions:

- BF3 DPU
- ConnectX7 NIC, based on the scale and performance considerations of a deployment.

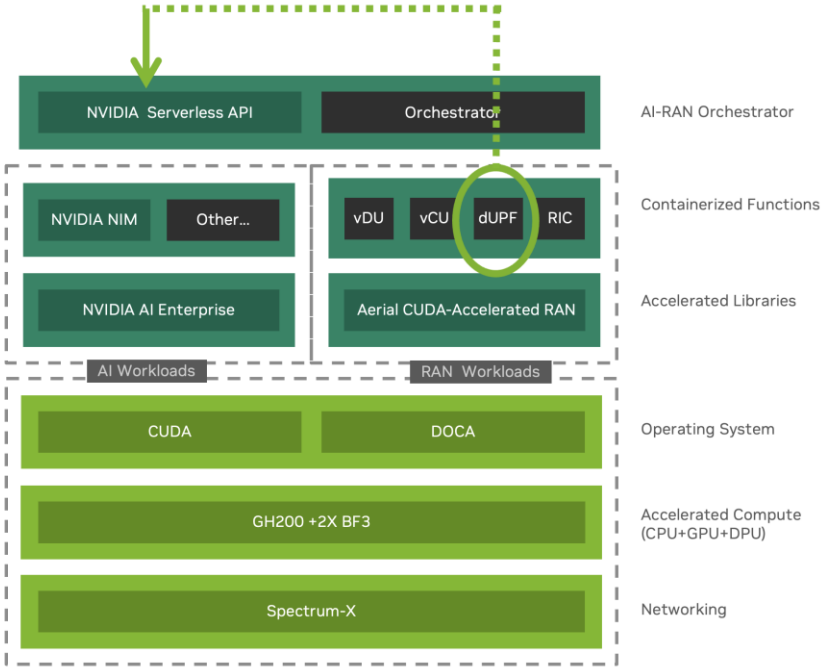
3.2 What is a software defined fronthaul interface?

The Network Interface Cards (NIC) used for AI-RAN deployments need to be capable of supporting high performance AI networking as well as supporting RAN timing and various RAN configurations. This software defined, multi-use interface enables easier network management and software upgradability, thereby supporting all three AI-RAN use cases described in section 1.1.

AI is increasingly getting integrated and used across all industry verticals. Applications are evolving rapidly to use Agentic AI, with capabilities to make decisions, adapt, and take actions on its own. It's designed to work with limited human supervision and can perform complex tasks. Applications such as autonomous vehicles, industrial robots, real-time language translation, XR camera – all require low latency and deterministic processing to deliver a high-quality experience.

Our vision for AI-RAN is to also process AI traffic transmitted over wireless networks by hosting 5G RAN and AI stack closely together in Telco Edge Datacenters - Aggregation Sites, Mobile Switching Offices (MSOs) and Central Offices (COs). This new architecture utilizes a 5G distributed User Plane Function (dUPF) to efficiently bridge the AI traffic to an AI inferencing stack with NVIDIA Serverless APIs and NVIDIA NIMs - Figure 6.

Figure 6. AI-RAN benefits for disaggregated network functions



AI-RAN therefore opens a new opportunity for Telcos – to convert their distributed Data Centers- MSOs, COs, and points of presence (POPs) into AI-RAN Data Centers and deliver

unique value by combining RAN and AI services to deliver low latency and deterministic experience for emerging new applications infused with AI that would otherwise be difficult to enable with today's centralized AI infrastructure.

This can only be done on an infrastructure that is carefully designed with the following components – MGX GH200 servers, Spectrum Switches and BF3 DPUs with their key capabilities as described. In the ensuing sections, we will examine the key networking considerations for the three AI-RAN use cases – AI for RAN, AI and RAN and AI on RAN.

3.3 What are the key networking considerations for AI-for-RAN?

There are many opportunities to utilize AI to improve spectral efficiency of RAN such as Channel Estimation/Prediction, Interference management, Beamforming, Deep Reinforcement Learning (DRL) based Modulation and Coding Scheme (MCS) selection and more. These can only be achieved with embedded accelerated hardware and software computing capability at Layer 1, that is fully programmable, as with NVIDIA CUDA accelerated libraries for radio signal processing, under the NVIDIA AI Aerial platform. If the L1 processing is performed by a fixed function accelerator, it is not possible to implement next generation AI-driven L1 optimizations. For instance, AI models for beamforming or dynamic spectrum sharing require continuous software updates and high-performance computing capabilities. Purpose-built RAN accelerators with integrated custom ASICs cannot support these emerging AI models, as they are designed for static functions and cannot accommodate iterative AI updates such as reinforcement learning.

A lot of progress is being made continuously on AI-for-RAN innovations, including some recent demonstrations endorsed by AI-RAN Alliance at Mobile World Congress (MWC) 2025. Notable public examples of these innovations include:

- SoftBank Demonstrates Performance Improvement in RAN Using AI with NVIDIA, Fujitsu
- Deepsig shows AI-Native Air Interface for 6G, using NVIDIA platforms
- Keysight, Samsung, NVIDIA Advance AI-For-RAN, using NVIDIA platforms

These innovations are early proof of the transformative gains possible in spectral efficiency. Purpose built RAN accelerators cannot support these continuous innovations as these accelerators are not programmable for the integration of such new techniques and are also outpacing multi-year cycles for developing custom hardware.

3.4 What are the key networking considerations for AI-and-RAN?

NVIDIA's NCP Telco Reference Architecture built with MGX GH200 servers and BF3 DPUs allows the 5G RAN, dUPF and the AI applications to be deployed on the same platform managed by Kubernetes. This brings tremendous TCO benefits as the platform resources (CPU, GPU, DPU/NIC) are dynamically allocated to RAN and AI functions thereby increasing their utilization and unlocking new AI monetization. NVIDIA's Spectrum-X, combined with BlueField-3 DPUs, optimizes AI-RAN performance by:

- Prioritization and Quality of Service (QoS): Leveraging AI-driven traffic management to prioritize latency-sensitive RAN traffic and ensure high-priority AI workloads.
- Increasing Bandwidth Utilization: From 50–60% to over 97%, speeding up data transfer for inference workloads.
- Reducing Latency: Advanced congestion control minimizes bottlenecks, ensuring real-time responsiveness.
- Improving GPU Utilization: Efficient network management maximizes GPU use for AI and RAN tasks. This includes software defined fronthaul.
- Lower inter-token latency: The increased bandwidth and optimized storage performance provided by Spectrum-X result in lower inter-token latency.
- Accelerated storage access: Spectrum-X improves read bandwidth by up to 48% and write bandwidth by up to 41% compared to traditional RoCE v2 protocols. This enhancement speeds up data retrieval and storage operations critical for inference tasks, particularly for techniques like retrieval-augmented generation (RAG).

Purpose built RAN accelerators and NICs lack these critical capabilities.

3.5 What are the key networking considerations for AI-on-RAN?

As enterprise applications integrate more AI capabilities and increasingly run on mobile networks, efficient processing of 'AI traffic' in distributed telco datacenters is critical to deliver the best quality and user experience. In this architecture, a dUPF is used to identify and bridge the AI traffic to the AI Inference software such as NVIDIA NIM. A purpose-built RAN accelerator card does not have the features and flexibility for an efficient dUPF (GTP tunnel encap/decap, Packet Classification, Receive Side Scaling (RSS), and QoS (Metering/Marking/Policing).

AI Agents are the next frontier for both consumer and enterprise applications. Agentic AI workloads require optimizations in the accelerated computing hardware and software stack, such that the compute latency for reasoning tasks is minimized.

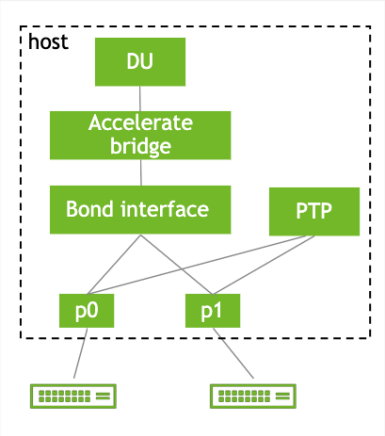
These Agentic AI optimizations are not possible with purpose-built RAN accelerators and NICs as these are not built for AI workloads.

3.6 How are fronthaul, midhaul, backhaul and AI traffic combined on a single NIC?

There are some important considerations when combining FrontHaul (FH), MidHaul (MH), BackHaul (BH) and AI Traffic on a single NIC such as a single BF3:

1. System Throughput Requirements – A single BF3 supports 2 ports of 200G throughput each. It is important to understand how much total traffic will be handled on the server between FH, MH/BH and AI applications. A typical deployment of up to 20 cells of 4T4R would utilize one port (200G) for FH and one port (200G) for MH/BH and AI traffic.
2. Fault-Tolerance, Redundancy and Service Assurance – Fault Tolerance and Redundancy for fronthaul is a key consideration for CoSPs. NVIDIA's NCP Telco Reference Architecture (RA) utilizes EVPN Multi-Homing with two BF3 NICs, with a port dedicated for RAN FH in each BF3 for robust fault tolerance and service assurance. As shown below, the FH ports on both NICs are configured Active-Active and present a single interface to the vDU application. If one of the port fails, the vDU application still maintains connectivity to the Radios – Figure 6.

Figure 3. Fault Tolerance and Redundancy in AI-RAN



3. Network performance degradation due to 'Accurate Scheduling' required for FH –

   Due to the synchronization and timing requirements of FH traffic, the NIC needs to use a special configuration namely Telecom Profile. This allows the NIC to do accurate scheduling of FH traffic with PTP timing synchronization. When combining MH/BH and AI traffic on the same NIC, there is a small performance degradation for the MH/BH and AI traffic. We estimate the performance degradation to be in the order of 10%. We believe this small degradation will not have a significant impact on the overall system performance.

4. C-RAN and D-RAN

4.1 Can AI-RAN be deployed in both C-RAN and D-RAN environments?

AI-RAN is a scalable software architecture that spans both C-RAN and D-RAN deployments. This scalability and software reuse is a key attribute and value proposition of AI-RAN.

NVIDIA recommends MGX GH200 and GB200 servers for Data Centers that host C-RAN. The modular MGX servers can be populated to support the power and cooling capacities from 2KW to 34KW or higher server racks in the Data Center. In future, many CoSPs are actively considering even higher density racks with Liquid Cooled (LC) technology.

For D-RAN, NVIDIA recommends MGX GH200 or MGX Grace C1 server with a PCIe attached GPU card such as L4 or L40S depending on 1) RAN capacity and coverage requirements and 2) AI and edge computing applications. This ensures that AI-RAN can meet the thermal, power and cost considerations in D-RAN deployments.

The Grace C1 server system can be designed into a Telco short-depth server with 250-300W total power consumption and withstand up to 55C outdoor temeparature.

4.2 Can D-RAN be deployed with a single NIC card?

A D-RAN deployment can use a single NIC card such as NVIDIA BF3 and CX7 with one port (200G) supporting FrontHaul (Radio Network) and one port (200G) supporting MidHaul/Backhaul and AI workload.

The actual deployment needs to carefully evaluate considerations such as – cell capacity, throughput, bandwidth, QoS for AI and RAN workloads, fan-out needs, fault-tolerance and redundancy requirements etc.

NVIDIA has announced the next generation ConnectX8 (CX8) NICs. When available in 2H-2025, CX8 can also be used for both C-RAN and D-RAN deployments.

5. References

[1] L. Kundu, X. Lin, R. Gadiyar, J-F. Lacasse and S. Chowdhury, "*AI-RAN: Transforming RAN with AI-driven Computing Infrastructure*", arXiv:2501.09007, Jan. 1, 2025

Notice

This document is provided for information purposes only and shall not be regarded as a warranty of a certain functionality, condition, or quality of a product. NVIDIA Corporation (“NVIDIA”) makes no representations or warranties, expressed or implied, as to the accuracy or completeness of the information contained in this document and assumes no responsibility for any errors contained herein. NVIDIA shall have no liability for the consequences or use of such information or for any infringement of patents or other rights of third parties that may result from its use. This document is not a commitment to develop, release, or deliver any Material (defined below), code, or functionality.

NVIDIA reserves the right to make corrections, modifications, enhancements, improvements, and any other changes to this document, at any time without notice.

Customer should obtain the latest relevant information before placing orders and should verify that such information is current and complete.

NVIDIA products are sold subject to the NVIDIA standard terms and conditions of sale supplied at the time of order acknowledgement, unless otherwise agreed in an individual sales agreement signed by authorized representatives of NVIDIA and customer (“Terms of Sale”). NVIDIA hereby expressly objects to applying any customer general terms and conditions with regards to the purchase of the NVIDIA product referenced in this document. No contractual obligations are formed either directly or indirectly by this document.

NVIDIA products are not designed, authorized, or warranted to be suitable for use in medical, military, aircraft, space, or life support equipment, nor in applications where failure or malfunction of the NVIDIA product can reasonably be expected to result in personal injury, death, or property or environmental damage. NVIDIA accepts no liability for inclusion and/or use of NVIDIA products in such equipment or applications and therefore such inclusion and/or use is at customer’s own risk.

NVIDIA makes no representation or warranty that products based on this document will be suitable for any specified use. Testing of all parameters of each product is not necessarily performed by NVIDIA. It is customer’s sole responsibility to evaluate and determine the applicability of any information contained in this document, ensure the product is suitable and fit for the application planned by customer, and perform the necessary testing for the application in order to avoid a default of the application or the product. Weaknesses in customer’s product designs may affect the quality and reliability of the NVIDIA product and may result in additional or different conditions and/or requirements beyond those contained in this document. NVIDIA accepts no liability related to any default, damage, costs, or problem which may be based on or attributable to: (i) the use of the NVIDIA product in any manner that is contrary to this document or (ii) customer product designs.

No license, either expressed or implied, is granted under any NVIDIA patent right, copyright, or other NVIDIA intellectual property right under this document. Information published by NVIDIA regarding third-party products or services does not constitute a license from NVIDIA to use such products or services or a warranty or endorsement thereof. Use of such information may require a license from a third party under the patents or other intellectual property rights of the third party, or a license from NVIDIA under the patents or other intellectual property rights of NVIDIA.

Reproduction of information in this document is permissible only if approved in advance by NVIDIA in writing, reproduced without alteration and in full compliance with all applicable export laws and regulations, and accompanied by all associated conditions, limitations, and notices.

THIS DOCUMENT AND ALL NVIDIA DESIGN SPECIFICATIONS, REFERENCE BOARDS, FILES, DRAWINGS, DIAGNOSTICS, LISTS, AND OTHER DOCUMENTS (TOGETHER AND SEPARATELY, “MATERIALS”) ARE BEING PROVIDED “AS IS.” NVIDIA MAKES NO WARRANTIES, EXPRESSED, IMPLIED, STATUTORY, OR OTHERWISE WITH RESPECT TO THE MATERIALS, AND EXPRESSLY DISCLAIMS ALL IMPLIED WARRANTIES OF NONINFRINGEMENT, MERCHANTABILITY, AND FITNESS FOR A PARTICULAR PURPOSE. TO THE EXTENT NOT PROHIBITED BY LAW, IN NO EVENT WILL NVIDIA BE LIABLE FOR ANY DAMAGES, INCLUDING WITHOUT LIMITATION ANY DIRECT, INDIRECT, SPECIAL, INCIDENTAL, PUNITIVE, OR CONSEQUENTIAL DAMAGES, HOWEVER CAUSED AND REGARDLESS OF THE THEORY OF LIABILITY, ARISING OUT OF ANY USE OF THIS DOCUMENT, EVEN IF NVIDIA HAS BEEN ADVISED OF THE POSSIBILITY OF SUCH DAMAGES. Notwithstanding any damages that customer might incur for any reason whatsoever, NVIDIA’s aggregate and cumulative liability towards customer for the products described herein shall be limited in accordance with the Terms of Sale for the product.

Trademarks

NVIDIA, the NVIDIA logo and NVIDIA Aerial are trademarks and/or registered trademarks of NVIDIA Corporation in the U.S. and other countries. Other company and product names may be trademarks of the respective companies with which they are associated.

VESA DisplayPort

DisplayPort and DisplayPort Compliance Logo, DisplayPort Compliance Logo for Dual-mode Sources, and DisplayPort Compliance Logo for Active Cables are trademarks owned by the Video Electronics Standards Association in the United States and other countries.

HDMI

HDMI, the HDMI logo, and High-Definition Multimedia Interface are trademarks or registered trademarks of HDMI Licensing LLC.

Arm

Arm, AMBA, and Arm Powered are registered trademarks of Arm Limited. Cortex, MPCore, and Mali are trademarks of Arm Limited. All other brands or product names are the property of their respective holders. "Arm" is used to represent Arm Holdings plc; its operating company Arm Limited; and the regional subsidiaries Arm Inc.; Arm KK; Arm Korea Limited.; Arm Taiwan Limited; Arm France SAS; Arm Consulting (Shanghai) Co. Ltd.; Arm Germany GmbH; Arm Embedded Technologies Pvt. Ltd.; Arm Norway, AS, and Arm Sweden AB.

OpenCL

OpenCL is a trademark of Apple Inc. used under license to the Khronos Group Inc.

Copyright

NVIDIA Corporation | 2788 San Tomas Expressway, Santa Clara, CA 95051
http://www.NVIDIA.com