



NVIDIA Aerial RAN Computer MGX 2U Short

A commercial AI-RAN deployment platform for software-defined, AI-native, 5G and 6G solutions.



NVIDIA Aerial RAN Computer MGX 2U Short

The NVIDIA Aerial™ RAN Computer (ARC) MGX 2U Short is a high-performance, energy-efficient, AI-RAN platform. Featuring NVIDIA Blackwell RTX PRO™ GPUs, it fuses advanced computing, connectivity, and sensing on the same infrastructure in telco-optimized form factors. The ARC MGX 2U Short is built for commercial distributed-RAN (D-RAN) deployment via network equipment providers (NEPs) and original equipment manufacturers (OEMs), enabling the transition from traditional RAN to AI-RAN and accelerating the path to AI-native 5G and 6G through software upgrades.

NVIDIA ARC MGX 2U Short is part of a larger NVIDIA ARC family that provides a range of AI-RAN deployment platforms for telecom operators through a wide ecosystem of OEM and original design manufacturer partners.

Optimized for Wireless Networks

A cost- and power-optimized platform for real-world telco deployments, ARC MGX 2U Short delivers the flexibility and performance to meet today's 5G requirements while providing a seamless path to AI-native 6G. Designed for both existing site upgrades and greenfield deployments, ARC MGX 2U Short highlights include:

- > **Compact, thermal-optimized**, and ruggedized form factor suitable for typical cell-site environmental conditions and cabinets
- > **Optimized power** that follows NVIDIA Max-Q principles, consuming ~300 watts of total system power at typical silicon 25C° ambient temperature
- > **Carrier-grade reliability** with telco-grade environmental compliance
- > **Deployability** into existing base stations or as a commercial off-the-shelf (COTS) server
- > **AI-RAN readiness**, supporting AI-for-RAN, AI-and-RAN, and AI-on-RAN

Built on NVIDIA Grace Blackwell

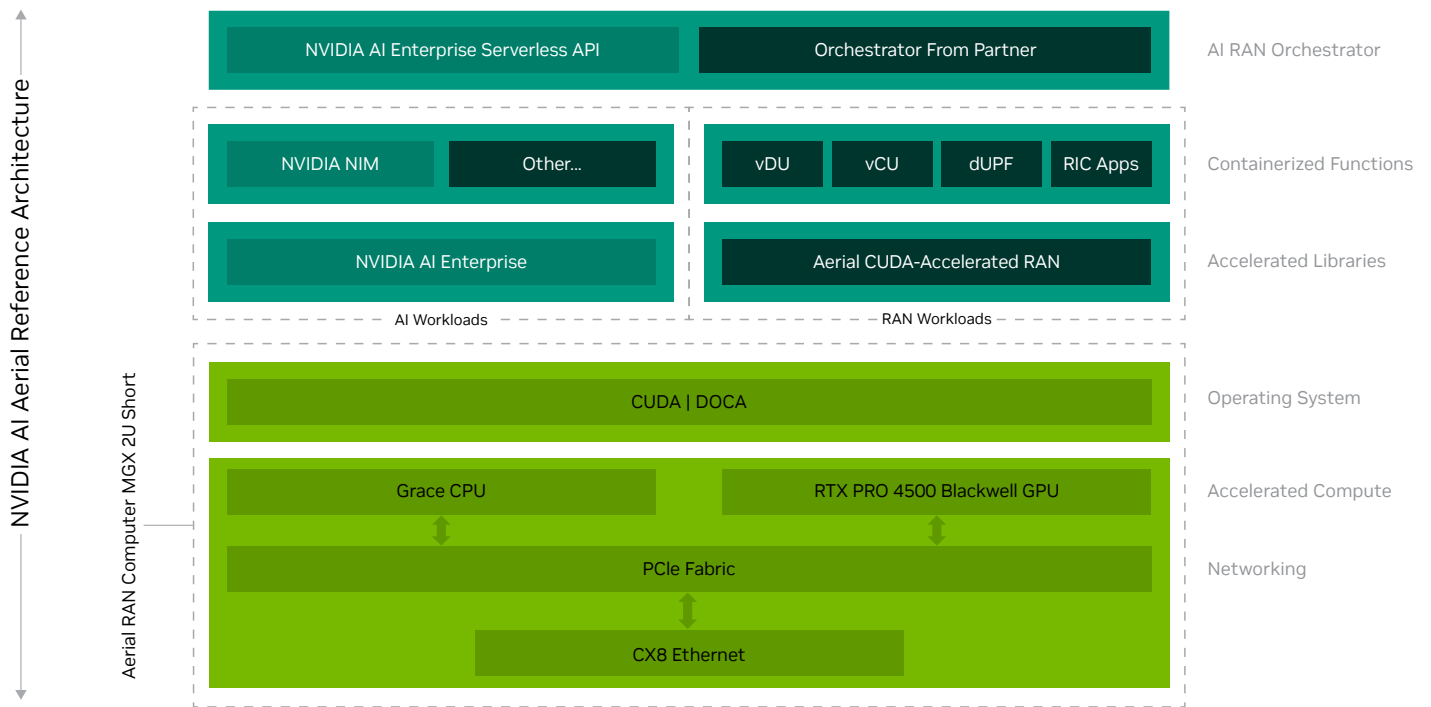
At the heart of NVIDIA ARC MGX 2U Short is the NVIDIA Blackwell RTX PRO 4500 BSE GPU with fifth-generation Tensor Cores and FP4 support, which powers real-time signal processing, AI acceleration, and up to 1.6 petaFLOP of AI compute. Built with NVIDIA-standard silicon, ARC MGX 2U Short leverages the NVIDIA Grace™ Blackwell architecture, including a high-performance NVIDIA Grace 72-core Arm® CPU to supercharge data processing and orchestration functions of the RAN stack.

Key Features

- > Powered by NVIDIA Blackwell RTX PRO 4500 GPUs, NVIDIA Grace 72-core Arm CPUs and NVIDIA ConnectX-8 NICs, with advanced security features
- > Cost- and power-optimized for cell sites, consuming ~300 W of system power, in a telco-grade form factor
- > Deployable into proprietary AI-RAN products by NEPs or COTS servers by OEMs
- > 5G FDD/TDD and massive MIMO today; ultra MIMO with software upgrade to 6G
- > Advanced AI/ML algorithms, edge AI applications and services, and sensor fusion
- > Homogeneous, software-defined architecture supporting D-RAN, C-RAN, and Open RAN

The reference design is highly optimized for delivering the scale and runtime performance required for AI-RAN networks. Key elements include:

- > **NVIDIA Blackwell-based RTX PRO 4500 BSE GPU** for RAN Layer 1, RAN Layer 2, and advanced distributed unit (DU) and centralized unit (CU) algorithms for higher spectral efficiency and performance
- > **NVIDIA Multi-Instance GPU (MIG) support** for dynamic multi-tenancy and utilization optimization during off-peak hours
- > **Inline GPU acceleration** of highly demanding RAN functions
- > **NVIDIA Grace CPU** with 72 Arm Neoverse V2 Cores, including Scalable Vector Extension 2 (SVE2) 4x128-bit single instruction, multiple data (SIMD) per core, to support RAN Layer 2, RAN Layer 3 and distributed user plane (UPF) and RAN intelligent controller (RIC) functions
- > **NVIDIA ConnectX™-8 Ethernet** for accurate timing and synchronization with direct radio unit (RU) connectivity, AI traffic routing, and advanced offloads.



NVIDIA Aerial RAN Computer MGX 2U Short, as part of AI Aerial reference architecture

Key Platform Security Features

- > Hardware root of trust with secure boot and firmware attestation across GPU, CPU, and network interface card (NIC)
- > Trusted execution environments (TEE) on the GPU to ensure secure model training and inference, even in multi-tenant deployments
- > NVIDIA Grace CPU security features, including secure boot, full memory encryption, and hardware virtualization
- > Encrypted PCIe connectivity between CPU, GPU, and NIC, protecting data movement across all platform elements
- > Inline cryptographic acceleration

AI Ready, 6G Ready

ARC MGX 2U Short supports current 5G requirements and is engineered for seamless 6G upgrades via a software-defined architecture. Its NVIDIA CUDA™ and Tensor Cores unlock advanced AI capabilities that can improve the radio network itself and support AI traffic growth while enabling new AI services on the same platform. Together, they power a high-performance AI-RAN platform that processes AI tokens alongside L1–L3 baseband functions in real time.

5G/6G RAN capabilities include:

- > **Support for current 5G standards**, including frequency-division duplexing (FDD) and time-division duplexing (TDD) multiple-input, multiple-output (MIMO) with 4 transceiver (TR) units, and TDD massive MIMO configurations such as 64TR
- > **Integrated high-speed networking** with precise timing interfaces for fronthaul, midhaul, and backhaul connectivity
- > **Full software stack compatibility** for virtualized RAN, Open RAN, and Cloud RAN architectures
- > **6G readiness**, including support for new spectrum bands, ultra MIMO (e.g., 256TR), and AI-enabled L1, L2, and L3 layers

AI capabilities include:

- > **AI-for-RAN** for using advanced AI and machine learning (AI/ML) algorithms and neural network models for RAN optimization, including massive MIMO, beamforming, channel prediction, power management, scheduling, and AI-powered dApps, rApps, and xApps
- > **AI-on-RAN** for hosting edge AI applications for low-latency inference and analytics
- > **Processing consumer-generated AI traffic** from chatbots, AI agents, generative-AI-based video and voice applications, AR glasses, and more
- > **Processing enterprise AI use cases** across robotics, smart cities, cybersecurity, video search and summarization, and computer vision solutions
- > **Sensor fusion** and integrated sensing and communication (ISAC) workloads

Multiple Deployment Options

ARC MGX 2U Short is built for commercial deployments and engineered for compatibility with existing cell sites. It fits into base stations already deployed in the field, delivering revolutionary technology through seamless upgrades. The same reference design can also be deployed on COTS servers or Cloud RAN with no performance compromise, ensuring forward scalability.

ARC MGX 2U Short enables a homogeneous, software-defined architecture across any deployment scenario, including:

- > **Proprietary AI-RAN products** by NEPs, integrated with NVIDIA ARC MGX 2U Short into their custom baseband units
- > **COTS-based AI-RAN servers** by OEMs, integrated with NVIDIA ARC MGX 2U Short
- > **Brownfield and greenfield deployments**, supporting both expansion of existing RAN systems or standalone installations

- > **Support for both distributed and centralized RAN** (D-RAN and C-RAN) scenarios
- > **Easy scalability** depending on workload and deployment scenarios, with a full range of options to support low- to high-density systems

Key Benefits for Telcos

Current networks aren't designed to process AI traffic. With NVIDIA-powered AI-RAN networks, telcos can efficiently serve growing AI traffic, introduce new AI services on the same infrastructure, and deliver significant improvements in performance, spectral efficiency, and energy efficiency via continuous software-defined innovation.

Key benefits for telcos include:

- > **Better customer experiences** for generative AI and agentic apps through low-latency edge inference and AI-enhanced network performance
- > **Improved spectral and energy efficiency** from AI-driven RAN optimization, delivering cost savings and better quality of service
- > **Faster innovation cycles and 6G readiness** via software upgrades, future-proofing investments for 6G and beyond
- > **Energy-efficient and cost-effective AI-RAN** deployments using merchant silicon-based COTS or proprietary systems
- > **Higher ROI** by running AI and RAN on the same distributed infrastructure, unlocking new AI services and increasing platform utilization
- > **TCO advantages** with continuity across hardware and software generations, using the CUDA ecosystem to advance innovation without application-specific integrated circuit (ASIC) lead times
- > **Easy installation** with a 2U short depth form factor that simplifies deployment and integration into existing telco infrastructure

With performance and efficiency already comparable to NEP custom ASICs and a continuous NVIDIA AI Aerial hardware-software roadmap, ARC MGX 2U Short is a game changer for the wireless industry, uniting AI compute and connectivity in a site-ready, energy-efficient, accelerated platform. It delivers AI-native 5G and 6G — enabling operators to innovate fast with a software-driven architecture.

Solution at a Glance

Technical Specifications

Architecture	NVIDIA Grace Blackwell
GPU	NVIDIA RTX PRO 4500 Blackwell Server Edition, with Multi-Instance GPU (MIG) technology; up to 2 MIG instances for each GPU
CPU	Grace-72 Core Neoverse V2
Ethernet	Up to 24 ports (up to 100 Gb/s per port), based on ConnectX-8 technology Dual 100 GbE / 200 GbE interfaces, TSN and eCPRI support
CUDA Cores	NVIDIA Blackwell generation
Tensor Cores	Fifth generation
Ray-Tracing Cores	Fourth generation
AI TOPS	1.6 PFLOPS of FP4 AI performance
Software Stack	NVIDIA AI Aerial libraries, CUDA, DOCA™, cuDNN™, TensorRT™ AI-RAN reference applications Third-party vRAN/O-RAN-compliant software
RAN Technology	Multi-RAT, upgradable to 6G, both FDD and TDD
RAN Performance	Standard MIMO (e.g., 4TR), massive MIMO (e.g., 64TR), ultra MIMO (e.g., 256TR)
Power Consumption and Energy Efficiency	~300 W (including fans and power supply), with typical silicon at 25 C° ambient temperature Gb/s/watt on par with traditional ASICs today
RAN Software Architecture	GPU-accelerated and software-defined L1 and L2
Environmental	Cell site (-5 C° to 55 C° ambient) Data center and enterprise (25 C° ambient)
Form Factor	Telco-grade short-depth 2U or 1U chassis, optimized for standard cell cabinets

Ready to Get Started?

To learn more about NVIDIA AI Aerial, visit:
developer.nvidia.com/aerial

© 2026 NVIDIA Corporation and affiliates. All rights reserved. NVIDIA, the NVIDIA logo, Aerial, ConnectX, cuDNN, CUDA, DOCA, Grace, RTX PRO, and TensorRT are trademarks and/or registered trademarks of NVIDIA Corporation and affiliates in the U.S. and other countries. Other company and product names may be trademarks of the respective owners with which they are associated. 5777400. OCT26

