



NVIDIA Aerial RAN Computer MGX 2U

A 2U platform for software-defined AI-RAN and AI inference.



NVIDIA Aerial RAN Computer MGX 2U server.

The NVIDIA Aerial™ RAN Computer (ARC) MGX 2U brings GPU-accelerated RAN and AI workloads onto a shared platform. Its modular design combines compute (both GPU and CPU) and networking for commercial deployments, from distributed cell sites to aggregated and centralized locations.

The platform supports current 5G RAN workloads and provides a software-defined path for developing 6G capabilities. It is part of a larger NVIDIA ARC family that includes a range of AI-RAN deployment platforms for telecom operators through a wide ecosystem of original equipment manufacturer (OEM) and original design manufacturer (ODM) partners.

Built for RAN and AI on Shared Infrastructure

ARC MGX 2U is designed to support virtualized RAN processing alongside edge AI inference. Operators can use AI for network optimization and to support applications such as video analytics, voice AI, and robotics near users and devices.

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 four 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 future 6G features, including new spectrum bands, AI-assisted radio processing, and higher-order MIMO

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

Key Features

- > NVIDIA MGX™ modular architecture in a 2U form factor
- > NVIDIA RTX PRO™ 6000 Blackwell Server Edition GPU acceleration
- > Configurable CPU options, including support for NVIDIA Vera and x86-based CPUs
- > NVIDIA ConnectX™-8 networking
- > Converged infrastructure for GPU-accelerated RAN and AI
- > Deployable into 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

- > **AI service delivery** by processing AI traffic from 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

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 NVIDIA CUDA™ ecosystem to advance innovation without application-specific integrated circuit (ASIC) lead times
- > **Easy installation** with a 2U form factor that simplifies deployment and integration into existing telco infrastructure

Technical Specifications

Architecture	NVIDIA MGX
GPU	NVIDIA RTX PRO 6000 Blackwell Server Edition, with configurations up to 2 GPUs; with Multi-Instance GPU (MIG) technology; up to 4 MIG instances for each GPU
CPU	Option 1: NVIDIA Vera CPUs Option 2: x86 CPUs
Ethernet	Up to 2 ConnectX-8 SuperNICs (1x 800GbE, 2x 400GbE) for East/West and fronthaul traffic; 1x NVIDIA BlueField™-3 (B3220) for North/South traffic
CUDA Cores	NVIDIA Blackwell generation
Tensor Cores	Fifth generation
Ray-Tracing Cores	Fourth generation
AI TOPS	4 PFLOPS of FP4 AI performance per GPU
Software Stack	NVIDIA AI Aerial libraries, CUDA, DOCA™, cuDNN™, TensorRT™ AI-RAN reference applications; DOCA dUPF 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	~1.4-2.5 KW based on workload and GPUs
RAN Software Architecture	GPU-accelerated and software-defined L1 and L2
Environmental	10–35°C (50–95°F) with no direct sunlight on the equipment (operational)
Form Factor	2U chassis

Ready to Get Started?

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

