GlobalLogic®
A Hitachi Group Company

Amplify AI-RAN Performance with NVIDIA GPU Acceleration

Benchmarks reveal significant network performance gains to advance AI-RAN and open new channels of monetization

A GlobalLogic thought leadership perspective

Table of contents

23% CAGR growth

Mobile data traffic projection for 2023 - 2030

*GSMA, The Mobile Economy 2024, 2024.

As demand for mobile connectivity soars, Radio Access Networks (RAN) face mounting pressure. In fact, mobile data traffic is projected to grow at a 23% CAGR from 2023 to 2030, reaching over 465 exabytes per month — a nearly fivefold increase*. Traditional RAN architectures, constrained by static resource allocation, vendor lock-in, and limited real-time adaptability, can no longer meet this scale.

Compounding the challenge are the rising complexities of 5G and forthcoming 6G deployments, the rapid expansion of edge computing and AI markets, and the emergence of new Telco business models, ranging from high-demand edge AI services and industrial/data-sovereign networks to enterprise network insights.

To meet these demands, forward-looking operators are exploring AI-powered RAN automation — not just to manage complexity, but to unlock dynamic optimization, energy efficiency, and adaptability at scale. From interference mitigation to predictive traffic control and multi-vendor orchestration, AI is reshaping the RAN landscape—and laying the groundwork for AI-native 6G.

What is AI-RAN?

AI-RAN is a Radio Access Network powered by Artificial Intelligence. AI is directly embedded into the fabric of RAN infrastructure. Designed to improve spectrum efficiency, reduce energy consumption, and adapt performance in real time, AI-RAN moves beyond static configurations toward continuous learning and autonomous network optimization. It leverages advanced AI models to enable this shift, in particular:

- **Machine learning** for adaptive resource allocation.
- **Deep reinforcement learning** for self-optimization.
- **Predictive analytics** for congestion avoidance.
- **Neural networks** for intelligent beamforming.

As mobile networks grow more complex, AI-RAN is poised to become a strategic enabler of dynamic, scalable, and intelligent connectivity. AI-RAN promises to help carriers enhance reliability and user experience, reduce costs, and unlock new monetization opportunities via:

Enhanced Spectral Efficiency	AI-driven signal processing could improve spectrum utilization, enabling higher throughput, better signal quality, and greater user capacity.
Edge Computing Enablement	Supports low-latency AI at the edge, accelerating applications like autonomous systems, smart cities, and industrial automation.
Smarter, Self-Adapting Networks	Enables real-time, autonomous optimization based on live traffic, user density, and environmental variables.
Automated Fault Detection and Self-Healing	Identifies anomalies, predicts failures, and resolves issues proactively to maintain service quality and reduce downtime.
Optimized Network Slicing	Tailors service quality and resource allocation for enterprise and IoT use cases, enabling differentiated offerings.
New Monetization Models	Unlocks value beyond RAN operations, allowing Communications Service Providers (CSPs) and vendors to run AI workloads on existing infrastructure and diversify revenue streams.

GlobalLogic AI-RAN and GenAI Pioneers

AI-RAN is still emerging, but few organizations bring the cross-disciplinary expertise needed to explore it meaningfully. At GlobalLogic, our foundation spans deep RAN development, cloud-native engineering, open network architectures, and AI innovation—positioning us to lead the next wave of wireless transformation.

We understand how to deploy and scale RAN more effectively, monetize network intelligence, and evolve legacy architectures into future-ready platforms.

We're not just enabling AI-RAN, we're redefining the very possibilities of what networks can become.

01

Foundations in radio access networks

We offer end-to-end RAN software expertise across L1–L3 stacks, developed through extensive work with LTE/5G nodes, small cells, CPEs, IoT/M2M devices, and more. Our AI/ML-powered analytics solutions already support adaptive, data-driven RAN behavior.

02

Cloud RAN engineering and lifecycle support

We design and implement cloud-native RAN architectures, including CI/CD automation, lifecycle management, testing, and security. Our work with ORAN specifications and SMO platforms helps accelerate the shift to modular, disaggregated networks.

GlobalLogic AI-RAN and GenAI Pioneers (cont.)

03

Open RAN integration and lab testing

With deep **ORAN** and **3GPP** knowledge, we handle conformance testing, deploy O-CU/O-DU/O-RU nodes, and optimize lab environments for validation and 5GC integration—enabling interoperability and system readiness.

04

Bridging AI-RAN and GPU acceleration

Leveraging NVIDIA GPU and CUDA expertise from industrial and automotive sectors, we now apply these capabilities to drive RAN innovation by:

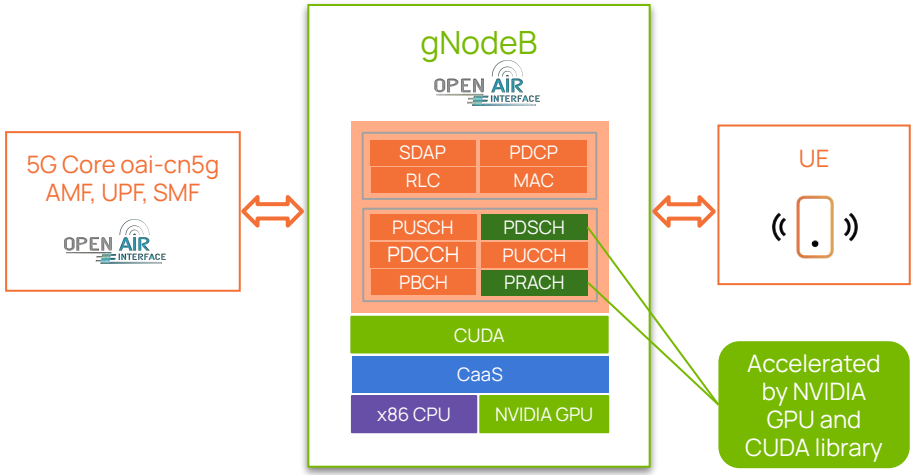
- Porting L1/L2 algorithms to CUDA
- Enabling GPU-accelerated baseband processing
- Building AI-enhanced radio signal workflows

This cross domain convergence empowers us to credibly prototype and scale next-gen AI-RAN solutions.

Proving the Concept:
AI-RAN Engineering in Action

Proof of Concept 01

5G Open RAN stack porting from CPU-based COTS to CPU + NVIDIA GPU-based COTS to evaluate performance and scalability



As a critical step toward demonstrating the real-world potential of AI-RAN, GlobalLogic conducted a targeted proofs of concept (PoCs) evaluations to assess GPU-based acceleration within 5G Open RAN environments.

The objective was to port essential components of a 5G Open RAN software stack from standard CPU-based COTS hardware to a hybrid CPU + NVIDIA GPU configuration, in order to assess performance and scalability gains. The study focused on two key physical channels:

1) **PRACH** (Physical Random Access Channel): Handles uplink user access and synchronization..
2) **PDSCH** (Physical Downlink Shared Channel): The primary channel for downlink data transmission.

From here the study concentrates on:

- LDPC (Low Density Parity Codes): The most computationally demanding function in the downlink chain.
- OFDMA (Orthogonal Frequency Division Multiple Access): A core component of PDSCH modulation and scheduling.

These channels were selected for their critical roles in 5G RAN operations and their high potential for GPU accelerated performance gains.

Outcomes: GPU Acceleration for PRACH and PDSCH Processing

01 PRACH Acceleration Results

The outcomes for the first PoC were incredibly promising.

Initial benefits appear at **mid-scale**, but the real value is unlocked in large, high-traffic environments where **scalability** and **efficiency gains are maximized.**

Scalability

GPU parallelism enabled efficient scaling across larger number of antennas with minimal overhead, demonstrating **strong horizontal scalability.**

This shows that GPU-accelerated virtualized Distributed Units (**vDUs**) running on COTS hardware **can handle increased traffic and computational loads effectively**. Performance bottlenecks are eliminated, and energy efficiency is significantly improved.

Operational Gains for the vDU

Increased UE capacity: Supports a higher number of simultaneous user equipment (UE) connections.

Reduced access latency: Dynamically scales gNodeB resources to minimize connection delays.

Improved energy efficiency: Delivers lower power consumption without compromising capacity.

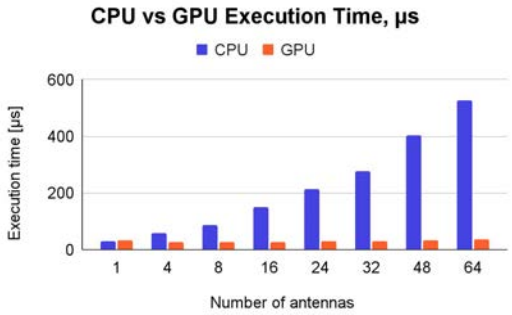
Enhanced UE detection: Increases reliability of user detection, resulting in a better overall user experience.

01 PRACH Acceleration Results (cont)

Performance Gains

With an enterprise-grade NVIDIA L40S GPU the processing time is **1500%** better with the GPU for 64 antennas.

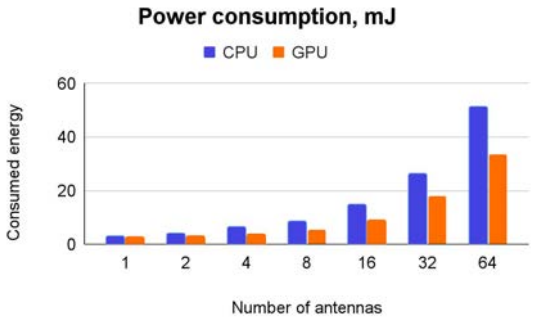
# of antennas	CPU [µs]	GPU [µs]
1	31	32
4	58	28
8	88	28
16	150	28
24	214	29
32	277	30
48	404	32
64	527	35



Power Consumption

The energy consumption of the GPU is consistently around **40% lower** for the same workloads.

# of antennas	CPU [mJ]	GPU [mJ]
1	2.96	2.45
2	4.09	2.82
4	6.35	3.72
8	8.65	5.37
16	14.82	8.75
32	26.41	17.91
64	51.06	33.33

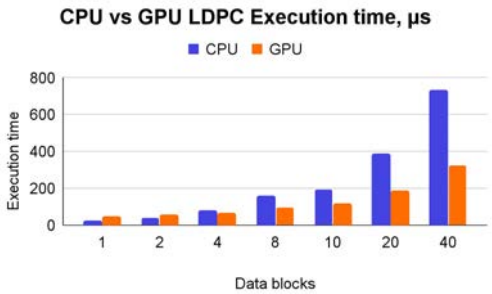


02 PDSCH Processing via LDPC Offloading

LDPC Acceleration

The team ported LDPC (Low-Density Parity Check) encoding - a key function in 5G downlink processing - to run on GPU.

Data blocks	CPU [µs]	GPU [µs]
1	21	46
2	40	51
4	78	65
8	156	95
10	193	115
20	382	184
40	727	326

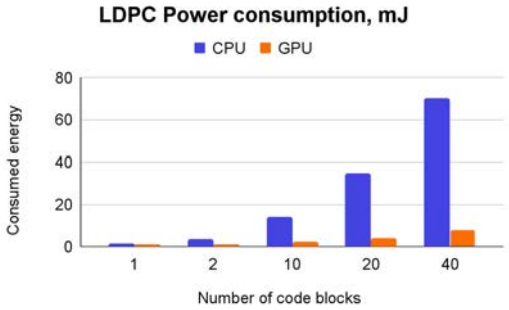


The Gains

Already with four blocks at the same time GPU executes faster and this gain grows up to **two times (200%)** for large loads (i.e. 40 data blocks) making GPU especially effective under high traffic loads.

The energy consumption of the GPU is up to **9 times (900%) lower** with number of data blocks growing.

Data blocks	CPU [mJ]	GPU [mJ]
1	1.24	0.67
2	3.17	0.79
10	13.7	1.87
20	34.21	3.55
40	69.89	7.75



02 PDSCH Processing via LDPC Offloading (cont)

Impacts on Large-Scale Deployments

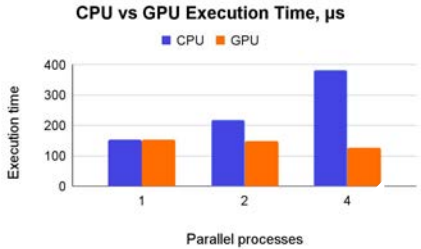
Higher throughput and user density: Delivers increased per-user throughput or supports more simultaneous UEs without compromising bandwidth.

CPU offloading with real-time performance: Offloads intensive workloads from CPUs while preserving real-time responsiveness.

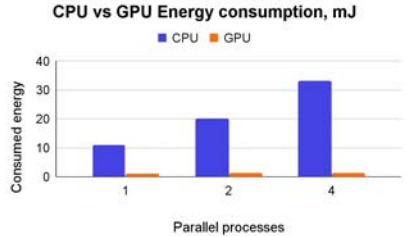
Optimized for high-traffic environments: Provides significant advantages in large gNodeB deployments under heavy load, where both computational performance and energy efficiency are essential.

03 OFDMA Processing in PDSCH

Parallel processes	CPU [µs]	GPU [µs]
1	151	151
2	217	146
4	380	125



Parallel processes	CPU [mJ]	GPU [mJ]
1	11	0.9
2	20	1.1
4	33	1.1



OFDMA acceleration

Orthogonal Frequency Division Multiple Access (OFDMA) is a key spectrum access technique that enables multiple users to share dedicated frequency subcarriers efficiently.

Performance testing revealed substantial gains when two or more parallel processes were used, resulting in a **33% to 67% reduction in execution time**.

Additionally, GPU-based processing demonstrated **12x to 30x lower energy consumption** compared to CPU-based execution, highlighting the significant efficiency advantages of GPU acceleration for OFDMA workloads.

Network impact

The computing efficiency gained through GPU acceleration frees up processing headroom, enabling the gNodeB to support more massive MIMO users without additional hardware.

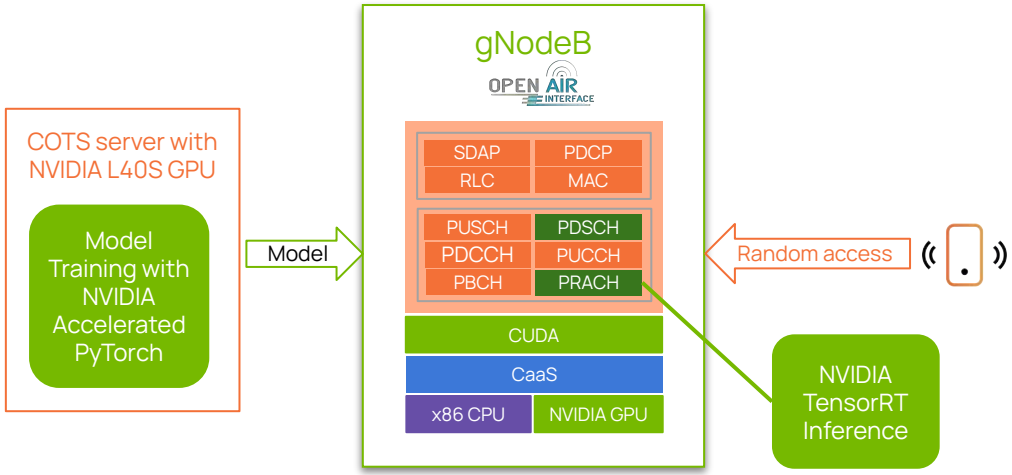
This spare compute capacity can be redirected to advanced signal processing or AI-based RAN intelligence, further enhancing network capabilities.

Moreover, GPUs deliver substantial power savings compared to CPUs, particularly in gNodeBs operating across wide frequency bandwidths and managing multiple cells simultaneously—making them ideal for dense, high-throughput deployments.

Proving the Concept: AI-RAN Engineering in Action

Proof of Concept 02

Reducing ghost preamble incidents via integrated deep learning



In our second proof of concept (PoC), we at **GlobalLogic** tackled a critical challenge in 5G RAN optimization — the ghost preamble phenomenon — a scenario in which base stations mistakenly interpret noise or interference as valid random access preambles from user devices.

These false positives result in wasted radio resources, increased signaling overhead, and reduced network performance, particularly in dense or noisy environments.

To mitigate this issue, we integrated deep learning models directly into the operational PRACH (Physical Random Access Channel) detection pipeline within a live 5G RAN stack. These models were trained using a combination of synthetic and real-world signal data, enabling real-time differentiation between genuine preambles and ghost signals. This resulted in improved access reliability and more efficient use of network resources.

PoC 2 Outcomes

Notably, this PoC also revealed promising & exciting opportunities for in-situ training of deep learning models directly within gNodeB infrastructure, paving the way for continuously adaptive, self-improving RAN systems representing a key enabler for **future AI-native networks**

Achieved 99.98% reduction in false positive detection, representing a major improvement over traditional threshold-based detection, where performance is on the level of 24.68% in challenging radio conditions.

Substantial reductions in false access attempts and unnecessary resource allocation from 75.38% to just 0.02%.

Improved network capacity, coverage, and user experience through more reliable access handling.

Demonstrated a modular, **operationally viable solution** with potential for commercial deployment.

Established a flexible framework combining specialized model architectures and adaptable deployment strategies.

Key Takeaways

PoCs demonstrate the potential benefits of replacing CPUs with GPUs for AI-RAN

Replacing CPUs with GPUs in AI-RAN environments unlocks several key advantages.

Performance Parity with Purpose-Built RAN

GPUs overcome compute bottlenecks of CPU-based Cloud RAN, enabling real-time signal processing at scale.

AI-Native Infrastructure

GPUs excel at accelerating deep learning, LLMs, and signal analytics, which are foundational for AI-driven RAN intelligence.

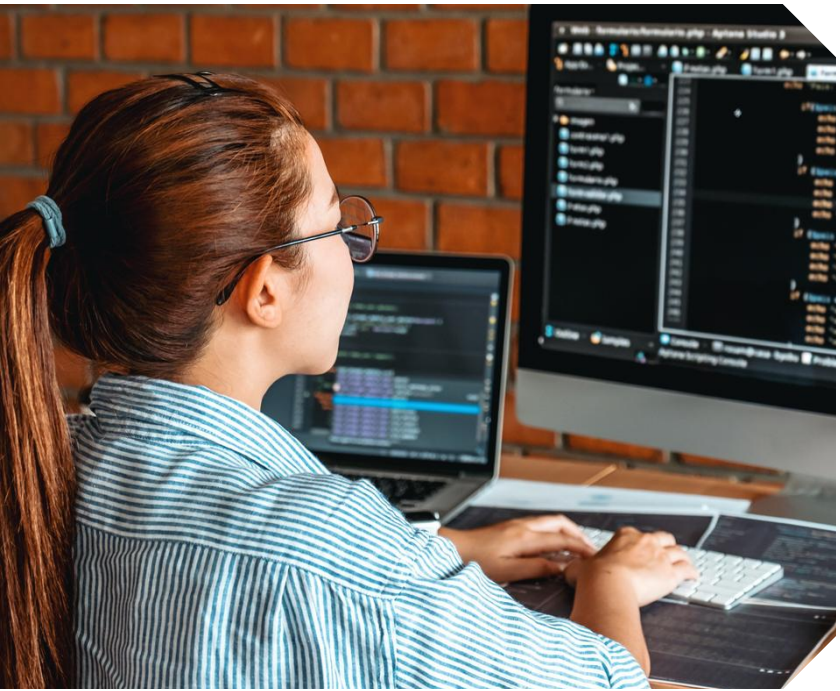
Greater Energy Efficiency

AI-optimized GPU processing reduces power consumption in dense traffic scenarios.

Converged Compute Platform

Supports both RAN and AI workloads on shared infrastructure, transforming networks into edge data centers and enabling new monetization models.

AI-RAN Can power intelligent, high-performance connectivity across key industry verticals



Railways

Supporting asset monitoring, operational safety, and real-time security systems.

Industrial

Enabling robotics, automation, IIoT connectivity, and energy-optimized logistics.

Smart Cities

Powering ADAS, transportation flow, infrastructure management, and traffic control.

Oil & Gas

Enhancing remote asset control, seismic data analysis at the edge, and worker safety.

These use cases unlock new value chains by turning RAN infrastructure into a programmable, revenue-generating platform.

The art of what's possible

Leading the charge to AI-RAN

We are at a pivotal shift in telecom infrastructure, one where AI and GPU acceleration converge to unlock unprecedented levels of network intelligence, scalability, and operational efficiency. Both GlobalLogic's proof-of-concept initiatives have clearly demonstrated that AI-RAN powered by GPU-based architectures is not just viable, it's transformative. From achieving up to 1500% faster processing in signal-heavy RAN tasks to reducing power consumption by up to 900%, the results signal not only technical feasibility but a compelling business case for accelerating AI-RAN adoption.

As mobile networks face mounting demand, latency sensitivity, and growing complexity, early movers in AI-RAN will be best positioned to lead. GlobalLogic brings together the critical capabilities—deep RAN expertise, advanced AI integration, and proven GPU engineering to empower Communications Service Providers (CSPs) and Network Equipment Providers (NEPs) evolve from connectivity providers to intelligent service enablers. By embracing AI-RAN today, organizations unlock immediate performance gains while laying the foundation for differentiated services, enhanced spectrum efficiency, and long-term monetization through AI-driven platforms. The opportunity is clear: AI-RAN is no longer just an innovation initiative, it is a strategic imperative for future-ready networks.

Authors and collaborations

Ganesh Seshadri, Vice President & General Manager, Network Providers, WiFi & Security BU

Ashay Punekar, Vice President, Communications & Networks BU

Serhiy Semenov, Director Engineering

Adam Radlinski, Solution Architect

Michal Celejewski, Software Developer

Mariusz Czapiewski, Software Developer

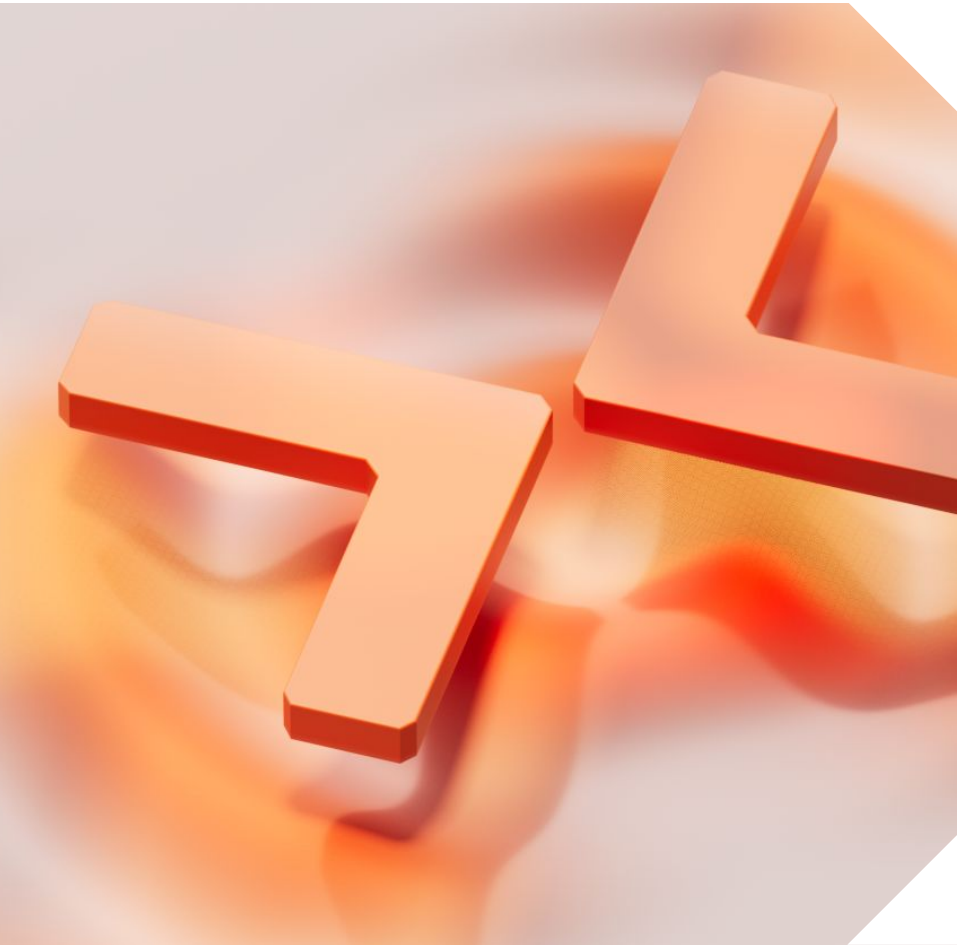
Mykhailo Morhal, Software Developer

Jakub Solich, Software Developer

Radoslaw Tereszczuk, AI Architect

Piotr Siminski, Software Developer

Piotr Przesmycki, Project Manager



Shaping the Future of Telecom

As AI-RAN evolves from concept to reality, **GlobalLogic stands at the forefront**—bridging deep telecom engineering with cutting-edge AI to unlock its full potential. From enabling GPU-accelerated RAN to pioneering AI-native network intelligence, we're not just anticipating the future of wireless we're actively building it.

The window of opportunity is now. Early adopters of AI-RAN won't just participate in the next wave of network evolution - they'll lead it.

GlobalLogic®
A Hitachi Group Company