

ARTICLE 05

AI-RAN and the Compute-Connectivity Economy

When the base station becomes an AI factory

AI-RAN and the Compute-Connectivity Economy

When the base station becomes an AI factory

AUTHOR Moderion Advisory

SERIES The Total Connectivity Economy

EVIDENCE CURRENT TO 6 September 2026

EXECUTIVE PROPOSITION

AI-RAN may turn distributed radio infrastructure into a shared compute platform, but technical feasibility is not yet proof of attractive utilization, isolation, energy or brownfield economics.

Executive summary

AI-RAN may turn distributed radio infrastructure into a shared compute platform, but technical feasibility is not yet proof of attractive utilization, isolation, energy or brownfield economics.

This evidence-led thought article develops an executive point of view using standards, official data, company disclosures and Moderion Advisory analysis. It separates established facts from emerging direction and reported implementation.

EXECUTIVE QUESTION

Under what operational and economic conditions can common infrastructure support both radio processing and AI workloads?

Key insights at a glance

- The AI-RAN Alliance distinguishes AI-for-RAN, AI-and-RAN and AI-on-RAN. These are different propositions: network optimization, shared infrastructure and third-party AI workload hosting should not be collapsed into one business case.
- The alliance reported 132 members and 33 demonstrations from more than 30 organizations around MWC 2026. Demonstrations indicate ecosystem activity, not standardized multi-vendor production economics.
- AI-for-RAN offers the clearest early operating cases: channel estimation, beamforming, scheduling, spectrum sensing, energy optimization and fault prediction.
- AI-and-RAN creates value only when hardware, orchestration and software can allocate resources predictably between time-critical radio workloads and variable AI demand without compromising service.
- AI-on-RAN requires local demand. Installed accelerator capacity is not an advantage if utilization is low, workloads cannot be placed securely or hyperscale alternatives remain cheaper and easier to consume.
- Portability is strategic. A tightly coupled accelerator and software environment may improve time to performance but can raise switching cost, constrain vendor choice and shift bargaining power.
- Accelerator comparisons are meaningful only against a specified radio configuration, traffic profile, latency target, software stack, precision mode and power boundary. Vendor peak-throughput claims are not interchangeable with sustained carrier workloads.

- Shared AI-and-RAN infrastructure requires hard reservation for time-critical radio demand, admission control for AI workloads and preemption or graceful degradation rules before utilization gains can be counted.
- A site power upgrade is a systems project involving grid feed, rectification, batteries, cooling, rack density, fire protection, structural loading and maintenance access, not simply an accelerator purchase.
- This article contributes to the Total Connectivity Economy by testing its enabling mechanism against universal reach, open interoperability, governed intelligence, trusted participation and sustainable economics.
- A March 2026 techno-economic preprint models attractive AI-RAN returns under selected traffic, demand, hardware and token-price scenarios. Because it is a preprint and scenario model, it is evidence of a testable investment hypothesis, not proof of production returns.
- A 2025 Communications Engineering O-RAN testbed demonstrates real system integration while also illustrating that testbed feasibility, interoperability and application latency remain far from a general carrier-scale economic case.

Exhibit 1 | The three AI-RAN propositions

Category	Primary value	Proof required	Principal risk
AI-for-RAN	Better network performance and cost	Measured KPI uplift versus baseline	Model drift and unsafe optimization
AI-and-RAN	Shared utilization of common compute	Isolation, scheduling and lifecycle TCO	Peak conflict and platform lock-in
AI-on-RAN	Local AI services and inference	Paid workload and placement advantage	Stranded capacity and weak demand

Note: Categories follow the AI-RAN Alliance; assessment criteria are Moderion Advisory's.

Exhibit 3 | Accelerator architecture decision matrix

Architecture	Strength	Constraint	Evidence required
Inline card	Deterministic data-path acceleration	Firmware and interface dependency	Tail latency, failover, portability
Lookaside	Selective offload and flexibility	Transfer and scheduling overhead	Sustained load and jitter
Integrated / custom SoC	High efficiency for stable functions	Long qualification and lower flexibility	Lifecycle workload stability
General accelerator	Broad AI software ecosystem	Reservation may strand capacity	Paid utilization after radio reserve

Note: Compare within a defined radio split and workload envelope; peak vendor specifications are not operating economics.

Core analysis

Three different propositions

AI for RAN uses models to optimize planning and operation. AI on RAN places AI workloads at network edge locations. AI with RAN allows radio and AI workloads to share accelerated infrastructure. These propositions are often blended into one story, although they have different buyers, architectures and investment cases. AI for RAN can create value without selling external compute. AI on RAN requires demand for distributed inference. AI with RAN adds the challenge of scheduling unlike workloads without compromising deterministic network performance.

The strategic promise is important. Operators already manage distributed sites, power, backhaul and trusted access. If accelerated resources can serve radio processing at peak periods and enterprise inference when spare, asset utilization could improve. Yet spare radio capacity does not automatically occur at the time, place and configuration demanded by AI workloads.

The economic gates

The first gate is utilization after reserving capacity for network resilience. The second is workload isolation: enterprise inference must not degrade radio performance or expose sensitive data. The third is energy, including accelerators, cooling and transport. The fourth is software portability across hardware and vendors. The fifth is lifecycle cost, particularly for brownfield estates whose refresh cycles and site constraints differ from cloud data centers.

An investable model must compare shared deployment with alternatives: centralized cloud, regional edge, dedicated enterprise infrastructure and conventional RAN. It should include incremental capex, site preparation, power, cooling, orchestration, model operations, security, support and stranded-asset risk. Revenue assumptions need identified workloads, contracted demand and realistic price erosion.

Architecture and control

Shared compute makes the orchestration layer strategically important. It must allocate resources across radio and AI workloads, enforce priority, measure consumption and support deterministic fallback. The software environment can become a new concentration point if models, accelerators and network functions depend on one proprietary stack. Multi-vendor openness therefore needs to be tested at runtime and operations level, not inferred from interface compliance alone.

The opportunity also extends beyond base-station sites. A compute-connectivity economy may combine central data centers, regional edge, far edge and devices. The optimal placement depends on latency, data gravity, sovereignty, model size, energy and utilization. The network's role is to make those placement choices observable and programmable.

A disciplined investment posture

Operators should treat AI-RAN as a staged portfolio of hypotheses. Begin with AI-for-RAN use cases that generate measurable network outcomes. Test shared infrastructure in controlled sites where enterprise demand and power conditions are credible. Preserve architectural options through modular hardware, portable software and transparent telemetry.

The sector should resist the idea that every base station becomes a commercial AI factory. Some sites may be valuable compute locations; many will not. The economic map must be derived from demand and constraints, not from the installed count of radio sites.

Moderion Advisory perspective

Separate the three investment cases

AI-for-RAN should be justified through radio and operating outcomes: coverage, capacity, energy, quality and engineering productivity. AI-and-RAN is an infrastructure utilization case requiring workload isolation, orchestration and a complete cost model. AI-on-RAN is a product-and-demand case requiring buyers, applications, placement advantage, service assurance and a channel.

This separation prevents savings from one category being used to justify speculative revenue in another. It also allows different ownership models: the operator may deploy AI-for-RAN as a network capability, procure shared infrastructure as a platform, and partner for application demand.

The compute-connectivity economic model

The model should calculate accelerator utilization by hour, radio reservation, peak coincidence, energy and cooling, software licensing, orchestration overhead, replacement cycle, support and security.

Revenue scenarios should be tied to committed workloads or a defensible distribution advantage rather than generic edge-compute forecasts.

The decision is location-specific. Dense metro or industrial sites with low-latency demand may support shared capacity. Rural sites designed for coverage may not. A common architecture does not imply a common deployment density, and the most efficient national portfolio may mix dedicated radio compute, shared platforms and centralized AI capacity.

Choosing an acceleration architecture

Inline acceleration places packet or signal processing directly in the data path and can reduce host work, but it raises dependency on card firmware, interfaces and failure behavior. Lookaside acceleration lets the host offload selected functions and may improve flexibility, while adding transfer and scheduling overhead. Integrated or custom SoCs can optimize performance per watt for stable workloads, but usually carry longer qualification cycles and less portability. General-purpose accelerators can support a broader AI ecosystem but may be economically weak when radio reservation leaves insufficient sellable capacity.

The comparison must use sustained throughput, tail latency, jitter, supported bands and carriers, failure recovery, software portability, power at representative load, rack density and lifecycle cost. A single watts-per-Gbps ratio is useful only inside the same workload definition; it should not be used to compare incompatible splits or feature sets.

A capacity model for co-located workloads

At interval t , schedulable AI capacity equals total qualified compute minus the radio safety reservation, platform overhead, resilience reserve and forecast-error buffer. AI demand may be admitted only when its execution window, data locality, security class and preemption tolerance fit the residual capacity. The financial model values completed, billable inference rather than nominal idle GPU hours.

The orchestrator needs a hierarchy: protect radio SLA first; preserve recovery capacity second; allocate contracted AI services third; admit interruptible workloads last. Forecast error should be charged to the business case through unused reserve or SLA risk. Peak and off-peak patterns must be measured at the actual cluster because radio, enterprise and event demand can coincide.

Power and thermal gating before deployment

The engineering baseline records utility capacity, rectifier efficiency, battery autonomy, cooling coefficient of performance, ambient-temperature envelope, rack and floor loading, backhaul, maintenance access and generator duty. The investment case then compares upgrade cost with avoided dedicated infrastructure and contracted AI demand. Sites that fail resilience or cooling gates remain radio-only or use centralized compute.

A portfolio decision is preferable to a universal architecture. Metro edge facilities, regional data centers and selected high-capacity sites may support shared compute, while coverage-led sites remain optimized for radio efficiency. Common orchestration can span these tiers without forcing identical hardware into every location.

Compute-connectivity convergence must create shared utility

AI-RAN contributes to the Total Connectivity Economy only when shared infrastructure improves productive utilization without compromising radio service, energy performance, portability or competition. The strategic test is not whether an accelerator can run two workload classes, but whether the integrated platform delivers reliable demand, defensible unit economics and governed allocation under real operating conditions.

Publication gate for AI-RAN economics

The peer-reviewed literature supports AI-assisted radio optimization and increasingly credible open-RAN experimentation. Evidence for profitable, shared AI-and-RAN infrastructure is earlier: reference architectures, demonstrations, simulations and a small number of techno-economic models dominate.

The investment thesis should pass only after a live cohort discloses reserved radio capacity, sellable AI capacity, workload utilization, power and cooling cost, performance isolation, software portability and cash contribution over a representative demand cycle.

Exhibit 2 | Deployment economics checklist

Dimension	Metric	Decision relevance
Utilization	Reserved radio + sellable AI hours	Determines productive asset use
Energy	kWh per radio / inference outcome	Reveals variable-cost advantage
Lifecycle	Hardware and software refresh	Captures obsolescence risk
Demand	Committed recurring workloads	Tests revenue credibility
Portability	Migration time and performance loss	Prices dependency

Note: Use site- and workload-specific data.

Executive action agenda

- Separate AI-for-RAN, AI-on-RAN and AI-with-RAN business cases and accountable owners.
- Model utilization by hour and location after reserving network capacity and resilience headroom.
- Measure total energy and cooling effects rather than accelerator efficiency in isolation.
- Test workload isolation, failover and software portability under production-like conditions.
- Scale only where enterprise demand or internal savings create a risk-adjusted return above the cost of capital.

Conclusion

AI-RAN is one of telecom's most consequential architectural options because it can connect network performance, distributed compute and enterprise AI. Its value will be proven not by the sophistication of a testbed, but by governed workload sharing, repeatable demand and superior lifecycle economics. In the wider series, this capability is treated as an enabler of the Total Connectivity Economy rather than as an end state in itself.

Selected evidence and references

- T014 | AI-RAN Alliance. AI-RAN Alliance mission and working areas. 2026. AI-RAN Alliance. <https://ai-ran.org/>
- T021 | Multiple authors. AI-Native O-RAN Architectures for 6G: Toward Real-Time Adaptation, Conflict Resolution, and Efficient Resource Management. 2026. IEEE Transactions on Network and Service Management. <https://ieeexplore.ieee.org/document/11483133/>
- T022 | Multiple authors. PACIFISTA: Conflict Evaluation and Management in Open RAN. 2025. IEEE Transactions on Mobile Computing. <https://ieeexplore.ieee.org/document/11007770/>
- T023 | Multiple authors. X5G: An Open, Programmable, Multi-Vendor, End-to-End, Private 5G O-RAN Testbed. 2025. IEEE Transactions on Mobile Computing. <https://ieeexplore.ieee.org/document/11039666>
- T035 | Multiple authors. A Survey on 5G Radio Access Network Energy Efficiency. 2022. IEEE Communications Surveys & Tutorials. <https://ieeexplore.ieee.org/document/9678321/>
- T036 | Multiple authors. Energy Efficiency in O-RAN Through Sleep Modes. 2025. IEEE ICC 2025. <https://ieeexplore.ieee.org/document/11160929/>
- T041 | Pettersson, Björkdahl & Holgersson. Profiting from AI: Evidence from Ericsson's Pursuit to Capture Value. 2025. California Management Review. <https://journals.sagepub.com/doi/10.1177/00081256251349140>
- T044 | Deloitte. 2026 Telecommunications Industry Outlook. 2026. Deloitte Insights. <https://www.deloitte.com/us/en/insights/industry/technology/technology-media-telecom-outlooks/telecommunications-industry-outlook.html>
- P002 | IEEE. AI-Native O-RAN Architectures for 6G: Toward Real-Time Adaptation, Conflict Resolution, and Efficient Resource Management. 2026. 10.1109/TNSM.2026.3684880. <https://ieeexplore.ieee.org/document/11483133/>
- P004 | IEEE. X5G: An Open, Programmable, Multi-Vendor, End-to-End, Private 5G O-RAN Testbed. 2025. 10.1109/TMC.2025.3580764. <https://ieeexplore.ieee.org/document/11039666>

- AI-RAN Alliance. AI-RAN Alliance White Paper. https://ai-ran.org/wp-content/uploads/2024/12/AI-RAN_Alliance_Whitepaper.pdf
- AI-RAN Alliance. AI-for-RAN White Paper. <https://ai-ran.org/documents/AI-RAN-WG1-AI-for-RAN-Whitepaper.pdf>
- AI-RAN Alliance. 2026 innovation and membership updates. <https://ai-ran.org/press-releases>
- ITU. ION-2030: enabling the intelligent optical network era. <https://www.itu.int/hub/2026/02/ion-2030-enabling-the-intelligent-optical-network-era/>
- AI-RAN Alliance. Platform and Infrastructure Orchestrator White Paper (2026). <https://ai-ran.org/wp-content/uploads/2026/02/AI-RAN-Alliance-Platform-and-Infrastructure-Orchestrator-White-Paper.pdf>
- AI-RAN Alliance. Publications and architecture documents. <https://ai-ran.org/publications/>
- PR014 | Kundu, L. et al. AI-RAN: Transforming RAN with AI-driven Computing Infrastructure (preprint, 2025). <https://arxiv.org/abs/2501.09007>
- PR015 | Gemmi, G., Polese, M. & Melodia, T. A Techno-Economic Framework for Cost Modeling and Revenue Opportunities in Open and Programmable AI-RAN (preprint, 2026). <https://arxiv.org/abs/2603.28680>
- PR016 | Hassouna, S. et al. Development of open radio access networks (O-RAN) for real-time robotic teleoperation. Communications Engineering 4, 176 (2025). <https://doi.org/10.1038/s44172-025-00524-0>
- PR017 | Implementing and Evaluating Security in O-RAN: Interfaces, Intelligence, and Platforms. IEEE Network 39(1), 227-234 (2025). <https://doi.org/10.1109/MNET.2024.3434419>