BRIEF

AI-RAN: Software-Defined Programmable Platform

Chetan Sharma

2026

chetan sharma CONSULTING

In collaboration with NOKIA

AI-RAN: Software-Defined Programmable Platform

Mobile operators are entering a period where the old capacity playbook is becoming harder to fund. Traffic continues to grow, uplink demand is becoming more important, densification remains slow and expensive, and AI-driven workloads will require tighter coordination between compute and connectivity. Additionally, the compute ecosystem is moving at a radically different cadence: AI models improve in weeks, GPU platforms advance in roughly two-year cycles, and enterprise demand for low-latency inference is beginning to collide with the network's historic role as a connectivity layer. The strategic question is no longer whether AI will influence the RAN, but where, when, and under what economic conditions operators should begin converging compute and connectivity.

The Architectural Logic

Three developments define why the moment has arrived. First, compute-network convergence: industrial automation, real-time vision systems, and low-latency AI applications — already in active operations, not future scenarios — require a unified fabric where compute and connectivity are jointly orchestrated rather than optimized as separate domains. AI-RAN provides that foundation, embedding AI natively into the radio access layer and transforming the RAN from a connectivity layer into the substrate for edge AI.

Second, AI-RAN changes the RAN innovation cycle. The important architectural shift is not dependence on a single GPU generation or even one vendor, but abstraction of RAN software from the underlying silicon — meaning RAN software runs on whatever compute the economics favor, not a locked hardware platform. The current validation path is GPU-led because that is where many AI-RAN gains are being tested, but the strategic value is portability: operators can match compute to site economics, refresh silicon as price-performance improves, and preserve the software estate across hardware transitions. If this abstraction holds in production, RAN algorithms move from fixed procurement-era capabilities to a faster, continuously improving software asset.

Third, AI-RAN creates an upside option beyond traditional RAN economics: selected non-RAN AI workloads may be hosted on the same distributed compute footprint where latency, data sovereignty, deterministic connectivity, or local processing create operator advantage. This should be treated as contingent upside until demand, pricing, utilization, and go-to-market models are proven.

The strategic case is to deploy AI-RAN first where the capacity, FWA, uplink, enterprise, and AI-hosting economics are strongest, then expand only as field data proves the power, reliability, and TCO envelopes.

A Three-Wave Roadmap — Sequencing Defines the Outcome

Nokia, T-Mobile, and NVIDIA have validated concurrent RAN and AI processing over-the-air at T-Mobile's AI-RAN Innovation Center in Bellevue — confirming that a shared GPU platform can handle both Layer 1 baseband and AI inference without compromising either. The algorithm portfolio unfolds in three waves, each building on the last. Wave 1 (active validation now, target commercial readiness Q4 2027) covers ML-assisted channel estimation, advanced MIMO receivers, RKHS channel-estimation techniques, and a Layer 2 portfolio targeting a favorable-site envelope of roughly 40–50% spectral-efficiency improvement under high-load, interference-limited Massive MIMO conditions; the realized

network-average gain will be lower and must be modeled by site typology. Wave 2 (late-2020s horizon) unlocks GPU-exclusive capabilities — ML-based beamforming, SRS channel prediction, DeepTx, built around large-scale matrix operations that are difficult to run on production ASIC hardware; their contribution is real but not yet quantified in field deployments. Wave 3 aligns with the AI-native 6G air interface 3GPP is standardizing (2030s horizon), with material capabilities arriving as software updates to hardware already deployed — though new spectrum bands will still require new radio hardware on any platform, and 6G's full compute requirements remain to be confirmed against GPU capabilities. The most important discipline is sequencing: Wave 1 is the appropriate scope for near-term capital decisions because it is the validated foundation and it generates the real-network training data, calibrated to an operator's own spectrum, antenna configuration, and traffic patterns that Wave 2 builds on. Nokia's engineering and R&D organizations project that successive hardware generations and full algorithm maturity could roughly double spectral efficiency relative to today's baseline — of which gains on the order of 20–25% have already been demonstrated in trials and simulations — targeting improvements of 90–100% toward a 2028 horizon, contingent on full maturation of the three-wave algorithm roadmap and under specific network conditions.

What the Gains Are Worth: The Spectral Efficiency Dividend

A spectral-efficiency gain is capacity earned on spectrum and sites the operator already owns. The practical value is not the percentage gain itself; it is the operator choice it creates - defer capital, monetize additional capacity, or improve experience in the sectors where the network is most constrained. Because more efficient encoding means fewer resource blocks carry the same traffic: PRB demand is inversely proportional to spectral efficiency. A 40% gain lets the same traffic occupy roughly 71% of the resource blocks it previously consumed. At constrained sectors, AI-RAN can deliver busy-hour relief that may defer some spectrum, densification, or site-split alternatives - without the timing uncertainty of auctions, clearing, or permitting. But demand does not pause. With major traffic categories growing in the mid-teens annually, and AI-era uplink traffic growing faster, a one-time efficiency gain behaves like a clock, not a permanent credit: on a capacity-bound sector, the window before re-congestion runs roughly fifteen to twenty-nine months. The dividend can be deployed in three ways: as deferred capital (avoiding near-term spectrum, densification, or site-split spend); as monetizable capacity growth, particularly in FWA markets where demand headroom remains material; or as improved user experience, disproportionately at the cell edge and uplink where AI-era traffic will press hardest.

Heterogeneous Networks - Right Compute, Right Site

Heterogeneity is the key mechanism that makes this an investment case rather than a wholesale platform bet. In a tier-1 macro estate, a relatively small share of sites typically carries a disproportionate share of total traffic - often dense urban locations, high-growth suburban sectors, FWA-constrained areas (markets where fixed-wireless demand exceeds current network capacity or competitive pricing pressure creates upgrade urgency), venues, enterprise zones, or transport corridors. Those are the places where the full Wave 1/2 accelerated-compute portfolio, the largest spectral-efficiency gains, and the most plausible AI Grid upside should be evaluated first. Lighter-load sites can participate through lower-power compute configurations running baseline ML-assisted functions, preserving architectural consistency and upgrade optionality without overprovisioning. Sites where the economics do not yet justify accelerated compute should remain on ASIC-enhanced platforms — purpose-built silicon that retains genuine advantages in power efficiency and operational simplicity and deserves rigorous comparison at sites where the AI-RAN economics are not yet compelling. The effect is a network tuned to its own traffic

distribution: capability and capital concentrate where the returns are highest, cost discipline holds everywhere else, and the upgrade path stays consistent across the estate.

The Stringent Performance Envelopes That Must Hold

Turning this case into an investment-grade proposition requires five conditions to hold, each verifiable with field data. Energy: Wave 1 algorithms operate at power thresholds broadly comparable to classical RAN, but the combined draw of RAN processing and AI inference at busy-hour production load has not yet been benchmarked live — until it is, any per-site energy figure in a TCO model should be treated as directional. Portability: the Hardware Abstraction Layer must hold across at least two GPU generations without material re-engineering at each transition, or the capital amortization model breaks down. Site typology: the 40–50% spectral-efficiency range is a best-case ceiling at the most demanding sites, not a network average — applying it uniformly will overstate revenue upside. Reliability: the platform must demonstrate equivalence with purpose-built ASIC on telecom-grade failure-mode and availability requirements, an area the GPU ecosystem is investing in but has not yet proven at scale. Operational complexity: AI-RAN must not impose net new burden on operators already running a classical RAN estate — running two software stacks and two validation cadences is net-positive only if that complexity is contained by the platform itself. The ASIC counter-case — purpose-built silicon with integrated NPU acceleration, lower power, and simpler operations — remains a genuine alternative that deserves rigorous comparison.

Investment Posture and Sequencing

The near-term investment case should be underwritten primarily by RAN economics: spectral efficiency, capacity relief, FWA headroom, and 6G readiness. AI Grid revenue should be treated as upside option value — contingent on operators proving customer demand, pricing, utilization, and go-to-market motion. The most defensible AI Grid opportunities will likely be workloads where the operator has structural advantages: low latency, data sovereignty, distributed geography, deterministic connectivity, and trusted local infrastructure.

AI-RAN is valuable because it lets operators concentrate programmable intelligence where network economics are most constrained, while preserving optionality elsewhere. AI-RAN is about building a network that compounds in capability with every silicon generation, generates AI-hosting revenue from infrastructure already in place, and is engineered to keep raising the spectral-efficiency ceiling well beyond the Wave 1 envelope as successive hardware generations and the broader algorithm roadmap mature. That potential is conditional: power and cost envelopes must hold in production, deployment must follow heterogeneous logic, and each wave of commitment must be grounded in field data. Operators who sequence correctly in 2026–2029 arrive at the 6G transition with years of accumulated training data and a software upgrade path that empowers optionality rather than giving it away.

This paper was commissioned by Nokia. Nokia provided access to field trial data and subject matter experts. All frameworks, analytical conclusions, and recommendations are those of Chetan Sharma Consulting and do not represent Nokia's official positions.