The Substrate Report

Volume 2

AI by Design: The Cognitive RAN

Architecture, Trust, and the Cross-Vertical Pattern

“AI is moving from application to architecture — and every network is starting to think.”

A T.A. Resendez Inc. publication
July 2026
taresendezinc.com · thesubstratereport.substack.com

The Substrate Report — Volume 2

AI by Design: The Cognitive RAN

Architecture, Trust, and the Cross-Vertical Pattern

A T.A. Resendez Inc. publication · July 2026

> *AI is moving from application to architecture — and every network is starting to think.*

How to read this Brief

Volume 2 of The Substrate Report is a three-episode arc. Each episode addresses one dimension of a single architectural thesis: AI has graduated from application running on top of the network to design center the network is built around. Episode 1 walks through the move inside telecom. Episode 2 walks through the trust architecture that makes it deployable at regulatory grade. Episode 3 walks through the cross-vertical pattern that proves the move is not telecom-specific.

This Brief consolidates all three episodes into a single document for readers who prefer the written form alongside the video podcast, or who want a printable reference for industry conversations. Each section corresponds to one episode briefing, with the same structure: executive summary, the story, the substrate-aware view, and what to watch in the next thirty days.

The Volume 1 wrap reference for callback continuity:

> *Volume 1, The New Industrial Stack, covered the spectrum war (Amazon's Globalstar acquisition), the AI layer at orbital speeds (AI-RAN, AI-by-Design at satellite scale), and the DePIN counter-movement. Volume 1's thesis: The next layer of the internet is being built in orbit, embedded in our streets, and crowdsourced by the world.*

Volume 2 picks up directly from the AI layer in V1 Episode 2 and walks it down to the ground.

Episode 1 — The Design Center Shift

When AI stops being an application and starts being the architecture.

Executive Summary

The AI-RAN Alliance, founded in 2024 by NVIDIA, SoftBank, and Samsung with roughly forty members today, is standardizing a fundamental shift in radio access network architecture: AI is no longer an application running on top of the network — it is the design center the network is built around. SoftBank's AITRAS orchestration platform is the most operationally advanced expression, turning the cell tower into an edge data center earning revenue from both mobile traffic and external AI inference. The architectural challenge is not bolting GPUs onto edge sites; it is co-designing hardware and software (programmable smart NICs, GPU Direct RDMA, in-line front-haul offload) so the data path actually flows. The same pattern is emerging simultaneously in automotive (Tesla's autonomy stack), energy (AI-orchestrated distributed energy resources), and edge infrastructure broadly. This is the architectural template the next decade of network deployment will run on.

The Story — AI Becomes the Design Center

The AI-RAN Alliance is what happens when silicon vendors, hyperscalers, and traditional telecom carriers agree that the conventional architecture is hitting a wall. Founded at MWC 2024, the consortium's charter is to standardize how AI workloads and cellular technology share the same underlying computing infrastructure. The alliance currently includes NVIDIA, SoftBank, Samsung, AWS, Arm, Ericsson, and Nokia among approximately forty members.

The operational proof point is SoftBank's AITRAS platform. AITRAS orchestrates a shared GPU pool at the cell tower base, running cellular workloads (neural channel coding, beam forming, signal decoding) alongside non-cellular AI workloads (generative AI inference, digital twin rendering, autonomous vehicle routing). When mobile traffic drops during off-peak hours, the orchestrator dynamically reallocates GPU resources to external AI customers. SoftBank reports hardware utilization approaching one hundred percent — a transformation from the historical pattern of custom ASICs sitting idle eighteen hours a day.

Mavenir's AI-by-Design framework approaches the same shift from cloud-native software, particularly in non-terrestrial network applications. The framework integrates AI at the design phase rather than retrofitted, pushing the architecture into Open RAN and satellite contexts.

The engineering reality, however, is harder than the marketing suggests. Bolting GPUs onto edge sites without redesigning the data path produces predictable failure: the host CPU saturates parsing Enhanced Common Public Radio Interface frames, managing PCI lanes, and extracting raw IQ samples from the radio. In field tests, CPU utilization pegs at one hundred percent while the GPU sits nearly idle. The architectural fix is programmable smart NICs (NVIDIA BlueField-3 is canonical) intercepting the fiber connection at the network card, processing the front-haul traffic on-card, and writing IQ samples directly into GPU memory via GPU Direct RDMA. The CPU is bypassed. The GPU runs hot. The architecture works.

The substrate-aware framing: This is the radio-access-layer expression of substrate-aware networking. The cognitive RAN — substrate-aware orchestration where the radio meets the

listener — is the same architectural pattern showing up across infrastructure industries simultaneously: automotive (silicon stacks designed around autonomy models, exemplified by Tesla's Drive Thor X and Foundation Model deployment), energy (AI-orchestrated distributed energy resources and virtual power plants), and edge broadly. AI-by-Design is not a vendor pitch; it is a universal architectural pattern.

The Substrate-Aware View

Three sectors are converging on the same architectural move at the same time. In each, AI moves from application to design center; the surrounding stack (radio, silicon, energy generation, transport) rotates around the AI workload rather than the other way around.

The unifying frame is **substrate-aware networks** — networks that know what kind of data they are carrying and adapt accordingly. The radio access layer, the energy resilience layer, the identity layer, the AI workload layer, and the transportation coordination layer all coordinate against the same substrate model. The network gains judgment, not just bandwidth.

For DePIN protocol builders, the implication is direct: the cell tower's transformation into an edge data center is the same pattern as a DePIN node's transformation into an AI inference site. Two revenue streams from one piece of hardware. For identity infrastructure, AI-RAN's trust requirements (every decision auditable, every model attested, EU AI Act compliance) telegraph where regulated deployments are headed. For Web3 infrastructure broadly, this is the substrate — the operating layer the next decade of deployments runs on.

What to Watch — Next 30 Days

1. **NVIDIA 6G Developer Program reference architecture releases** — specifically whether hardware-software co-design becomes published reference, not just compute reference. If so, AI-by-Design crosses from vendor pitch to industry standard.
2. **AI-RAN Alliance white paper publications** — the alliance's multi-vendor workload coordination specifications are the operational test of whether forty members can actually agree on standards rather than competing toolchains.
3. **Mavenir + SATELLITE 2026 / Mobile World Congress carryover** — Mavenir's AI-by-Design positioning against Ericsson's legacy silicon discipline is the most visible architectural tension in telecom this quarter.
4. **Tesla Foundation Model V2 indicators** — Tesla's automotive AI stack iterations telegraph the cross-vertical pattern's velocity. V2 releases would confirm the design-center shift is generational, not cyclical.
5. **EU AI Act enforcement actions on critical infrastructure AI** — first enforcement cases will define what "high-risk AI" auditing actually requires in practice. Until enforcement, the regulatory framework is theoretical.

Episode 2 — Trust at the Edge

How you certify AI decisions on a 10ms budget without breaking the network.

Executive Summary

The EU AI Act classifies critical infrastructure AI as high-risk and mandates audit trails plus explainability for every decision. AI-RAN deployments make decisions on a ten-millisecond Near-RT RIC control loop budget — a budget that cannot accommodate live explainability computation. The trap is binary: running XAI in the live inference loop crashes the network on latency, while skipping XAI fails the regulatory audit. The architectural fix is to move XAI out of the live inference path entirely, into the asynchronous federated learning pipeline at the Non-RT RIC, where confidence thresholds gate every model update before it is incorporated into the global model. The global model becomes inherently trustworthy without adding a microsecond to the live loop. Channel distribution shifts add a separate challenge — neural channel estimators trained on idealized noise collapse in real-world Rayleigh fading and Doppler conditions. The three-layer resilience stack (LoRA channel adapters, Network Digital Twins for synthetic stress testing, 3GPP RRC fallback to OFDM and LDPC) makes the architecture survive contact with reality. This is the trust frame substrate-aware networks were built for.

The Story — The Trust Architecture

The first AI-RAN trust problem is regulatory. The EU AI Act, in force as Regulation 2024/1689 and now in phased enforcement for critical infrastructure, classifies AI deployed in URLLC contexts (autonomous driving networks, smart grids, industrial robotics, emergency dispatch) as high-risk. Audit trails are mandatory. Explainability of individual decisions is mandatory. The regulator's first question when something goes wrong will be: why did the network make that decision? Operators that cannot answer fail the audit. Operators that try to answer using live-loop explainability fail on latency.

The Near-RT RIC operates on a strict ten-millisecond control loop budget per O-RAN Alliance specifications. Live computation of Shapley values for every packet schedule, every modulation choice, every resource allocation decision would exceed the budget many times over. The architectural escape from this trap is to recognize that explainability does not have to live in the same timescale as inference. Local Open Central Units train models on local cell-site traffic, retaining all raw data on-site and transmitting only updated mathematical weights to the Non-Real-Time RIC over the A1 interface. The Non-RT RIC applies XAI confidence thresholds — RuleFit translating weight updates into human-readable decision rules, SHAP providing feature attributions tied to specific RF parameters — to each update before incorporation. Updates that fail the confidence threshold are not aggregated. The global model pushed back down to cell sites is therefore inherently trustworthy at the regulatory grade, and the live inference loop at the Near-RT RIC operates on its ten-millisecond budget undisturbed.

The second AI-RAN trust problem is environmental. Neural channel estimators are typically trained on idealized Additive White Gaussian Noise. In real deployments — a high-speed train

introducing severe Doppler shift, a thunderstorm producing complex multipath fading, a new building changing the local reflection geometry — the model encounters distributions it has never seen. The confidence matrix collapses and the connection drops. The architectural answer is a three-layer resilience stack: LoRA-style channel adapters update approximately 3.5 percent of the model's parameters on the fly to recalibrate to local conditions; Network Digital Twins ingest site-specific 3D geometry and ray-tracing data for synthetic stress testing before the model touches live spectrum; and 3GPP RRC capability negotiation provides hard fallback to traditional OFDM and LDPC coding when AI confidence collapses. The architecture is explainable in async, adaptable in real time, and gracefully degradable when reality breaks the model.

The substrate-aware framing: The cognitive RAN is substrate-aware trust orchestration where the radio meets the listener. The network knows what kind of data it is carrying and therefore knows what trust attestation that data requires. The identity layer carries the attestation chain. The federated learning pipeline carries the model lineage. The 3GPP fallback layer carries the deterministic safety net. Three layers, three timescales (milliseconds for fallback, hours for federated aggregation, days for audit chain), one substrate-aware operating model. The same architectural pattern applies in automotive (every Tesla steering decision must be auditable, but operates on millisecond budgets) and in energy (grid-balancing AI must be explainable to utility regulators, but operates on sub-second budgets). The trust architecture is not telecom-specific; it is substrate-aware-specific.

The Substrate-Aware View

Substrate-aware architectures are not faster than substrate-blind architectures. They are more honest about what data is moving through them and what each piece of data requires for safe handling. Trust ceases to be a feature bolted onto the network and becomes a property of the substrate.

For AI-RAN, this means trust is woven into the federated learning pipeline, the identity attestation chain, and the 3GPP safety net simultaneously. For automotive AI-by-Design, it means trust is woven into the Tesla Foundation Model's training and attestation infrastructure, not retrofitted at deployment. For AI-orchestrated distributed energy resources, it means trust is woven into the model attestation chain at the inverter / generation / load-management level, not added after grid disturbances reveal opacity.

For DePIN ecosystems building AI inference capacity, the trust architecture is portable. Local nodes train models, send only weights to coordinators, coordinators apply XAI thresholds before aggregating. Same pattern, different substrate. For DID infrastructure (Hedera, ENS, DIF-aligned protocols, Sovrin), the cryptographically verifiable attestation chain that AI-RAN will require is exactly what these protocols have been building in research mode. The deployment timeline is shorter than most teams realize.

What to Watch — Next 30 Days

1. **First EU AI Act enforcement action involving critical infrastructure AI.** Pilot enforcement cases will define what "high-risk infrastructure AI audit trail" actually requires in practice. The published decision will shape architecture for the next three years.
2. **NVIDIA Aerial CUDA-Accelerated RAN 2026 release.** Specifically whether the release includes reference implementations of federated learning + Non-RT RIC architecture, or only raw GPU compute reference. Reference implementations would cross the architecture from vendor pitch to industry standard.
3. **O-RAN Alliance xApp / dApp framework specifications for AI inference.** Watch for the next specification release defining how AI inference modules deploy on the Near-RT RIC with audit trail metadata attached. Spec maturity here is the unblock for production AI-RAN deployments in regulated markets.
4. **3GPP Release 19 production deployments referencing AI/ML integration.** Production deployment timing tells us whether 3GPP standardization is on the AI-RAN timeline or trailing by a year.
5. **Tesla Foundation Model V2 indicators regarding attestation infrastructure.** Tesla's roadmap on auditable AI decisions in regulated automotive contexts is the cross-vertical reference point for whether the substrate-aware trust architecture works at scale.

Episode 3 — AI by Design, Beyond Telecom

The cross-vertical pattern — and the six-organ anatomy of the substrate-aware network.

Executive Summary

The AI-by-Design pattern that Volume 2 opened with in telecom is not a telecom phenomenon. The same architectural move — AI graduating from application running on top of the infrastructure to design center that the infrastructure is rebuilt around — is occurring simultaneously in automotive, in distributed energy, and in enterprise private networking. Tesla's Robocar stack puts a Vision Language Action multimodal model at the silicon design center, with NVIDIA Drive Thor X providing 800+ safety-capable cores and the Tesla Foundation Model V0/V1 framing the autonomy decision layer; the steering wheel literally folds away when the AI is in control. AI-orchestrated distributed energy resources move AI from analytics overlay to native grid infrastructure, with legacy SCADA control becoming the safety net underneath rather than the design center on top. NTT Data Managed Private 5G plus integrated IT/OT security demonstrates that enterprise private networks are now deployed AI-native from day one, security woven into the substrate. The substrate-aware network is the operating layer that orchestrates across all four sectors. Six functional organs — DePIN, DID, DeSci, DeFi, DPAi, DAOs — plus vCons as the nervous system constitute the anatomy that inhabits telecom, automotive, energy, and enterprise simultaneously.

The Story — The Cross-Vertical Pattern

Automotive: Tesla's Robocar and the Foundation-Model Silicon Stack

Tesla's Robocar stack is the automotive expression of AI-by-Design. The silicon at the center is NVIDIA's Drive Thor X — over 800 safety-capable cores architected specifically to host multimodal generative AI inference on a moving platform. Tesla's Foundation Model V0 and V1 frame the decision layer; the Vision Language Action multimodal model fuses camera, microphone, LiDAR, and radar input into language-grounded action sequences. The OpenXLA toolchain compiles the model down to Drive Thor X with the efficiency the autonomy budget requires. The vehicle's sensor stack — five terabit cameras plus one LiDAR plus one radar plus twelve microphones — feeds the inference engine continuously. The foldable steering wheel is the visual signature of the design-center shift: when the AI is in control, the human interface retracts. The car is a multimodal AI inference platform that happens to have wheels.

This is the inverted architecture. In the prior generation, the silicon was designed around the radio (telecom) or around the powertrain control unit (automotive). In this generation, the silicon is designed around the model, and the radio or the powertrain rotates into a peripheral role. The Tesla deployment is the most operationally visible instance.

Energy: AI-Orchestrated Distributed Energy Resources

For decades the energy industry deployed AI as an analytics overlay sitting on top of legacy SCADA grid control — predictive maintenance recommendations, demand forecasting, anomaly detection. The grid still operated on deterministic control loops; AI advised. The current architectural move flips that: AI becomes the orchestrator of the grid as native infrastructure, with legacy control becoming the safety net underneath rather than the design center on top.

Distributed energy resources — rooftop solar, behind-the-meter batteries, electric vehicle chargers, smart thermostats — are no longer endpoints reporting to a central utility. They are participants in an AI-orchestrated dispatch network. The AI predicts local generation, forecasts local demand, optimizes the dispatch path through the grid, and arbitrages against the wholesale market in real time. The utility's role becomes attestation of safety and regulatory compliance rather than dispatch authority. The same EU AI Act and audit-trail requirements that Episode 2 mapped onto AI-RAN apply here directly; the trust architecture is portable.

Enterprise: NTT Data Managed Private 5G plus IT/OT Security

NTT Data's Managed Private 5G offering is the enterprise expression of AI-by-Design. Industrial campuses — factories, ports, warehouses, refineries — deploy private 5G networks with AI integrated at the design phase, not retrofitted at deployment. Security is woven into the substrate: IT/OT integration is not a separate workstream but a property of how the private network is architected. The same Mavenir MoVerge portfolio and Open RAN specialization that surfaced in Episode 1 applies in enterprise contexts as well — the cloud-native architectural assumption is now the default, not the variant.

The Cross-Vertical Frame

Four sectors. Telecom (AI-RAN cell tower becomes edge data center). Automotive (Drive Thor X and Foundation Model V0/V1 as silicon design center). Energy (AI-orchestrated distributed energy resources as native grid infrastructure). Enterprise (NTT Data Managed Private 5G with AI-native security). The same architectural move in the same quarter. AI-by-Design is universal architectural pattern, not a vendor pitch.

The Substrate-Aware View — The Six-Organ Anatomy

The substrate-aware network is the operating layer that orchestrates across all four sectors. The six-organ anatomy frames how the substrate operates:

- **DePIN** — Physical Substrate. Decentralized Physical Infrastructure Networks providing the bedrock compute, storage, and connectivity layer. Across the four sectors: AI-RAN cell tower (telecom), Tesla vehicle fleet (automotive), distributed energy resource hardware (energy), private 5G base stations on enterprise campuses (enterprise).
- **DID** — Identity and Trust Substrate. Decentralized identifiers and cryptographic attestation for every node and every decision. Across the four sectors: cell tower identity attestation (telecom), vehicle identity and over-the-air model attestation (automotive), grid asset identity and dispatch attestation (energy), enterprise device identity and IT/OT trust attestation (enterprise).
- **DeSci** — Knowledge Substrate. Decentralized scientific verification, model lineage, peer review of inference behavior. The independent auditor of the substrate-aware network. Maps directly onto EU AI Act audit-trail requirements.
- **DeFi** — Economic Substrate. The settlement and market layer underneath the orchestration. Across the four sectors: AI-RAN compute marketplace (telecom), vehicle-to-grid energy market (automotive crossing into energy), real-time DER market (energy), private 5G usage-based billing (enterprise).
- **DPAi** — Physical AI / Real-World Learning. Where the AI meets the physical substrate it is shaping. Tesla Foundation Model V0/V1 (automotive). Channel adaptation models for AI-RAN (telecom). DER orchestration AI (energy). Industrial AI on private 5G (enterprise).
- **DAOs** — Governance / Collective Decision-Making. The decision-making and coordination layer for the substrate community.

vCons cross-cuts all six as the Neural Substrate of Web3. vCons is the nervous system that carries the consent, context, and conversation state across the six organs. JLINC consent at the protocol level. Every decision-carrying conversation between humans and substrate, between substrate components, between regulatory layer and operational layer travels on vCons.

The substrate-aware bridge line — *"The cognitive RAN is the radio-access-layer expression of substrate-aware networking. AI-by-Design at the RAN is substrate-aware orchestration where the radio meets the listener"* — generalizes. The cognitive vehicle is the automotive-layer expression. The cognitive grid is the energy-layer expression. The cognitive enterprise is the campus-layer expression. AI-by-Design across all four is substrate-aware orchestration meeting the listener at four different layers.

What to Watch — Next 30 Days

1. **Tesla Foundation Model V2 release indicators.** Tesla's roadmap on the next-generation Foundation Model directly determines the velocity of the cross-vertical pattern in automotive. V2 with broader regulatory attestation infrastructure would confirm the design-center shift is generational, not cyclical.
2. **Mavenir MoVerge / NTT Data enterprise deployments at scale.** Public-customer-named Managed Private 5G deployments at named industrial campuses telegraph whether the enterprise expression is moving from pilot to standard procurement. Particularly watch automotive manufacturing, semiconductor fabrication, and pharmaceutical campuses.
3. **AI-orchestrated DER market emergence in the U.S.** California, Texas, and New York are the leading regulatory venues for AI-orchestrated DER deployment. Watch for utility tariff filings that explicitly reference AI dispatch and attestation requirements. The first tariff that requires AI-native trust attestation moves the architecture from market-edge to market-default.
4. **The first cross-vertical architecture reference.** Watch for a vendor — likely NVIDIA, possibly Mavenir, possibly an industrial automation company — publishing a reference architecture that explicitly spans telecom AI-RAN, automotive Drive Thor X, and DER orchestration on a unified substrate. The publication itself is the proof point that the four sectors are converging on one architectural model.
5. **vCons community standardization momentum.** vCons IETF / standards-body progression is the test of whether the nervous-system layer is ready to carry the substrate-aware operating model at the scale the four sectors will demand. Particularly watch for cross-industry adoption signals beyond telecom.

Volume 2 Wrap

Episode 1 walked through the design-center shift in telecom — AI moving from application to architecture, cell tower becoming edge data center, the AI-RAN Alliance standardizing the move, the smart NIC and GPU Direct RDMA architectural fix that makes the data path work. Episode 2 walked through the trust architecture — XAI moved out of the live 10ms inference loop into the asynchronous federated learning pipeline, the three-layer resilience stack handling channel distribution shifts, the substrate-aware trust frame the architecture was built for. Episode 3 closed by mapping the same architectural move across automotive, energy, and enterprise, and showing why the substrate-aware operating layer was always the right frame — not because it was telecom-specific, but because the cross-vertical pattern requires it.

Three episodes. One architectural thesis. Volume 2 closes here.

The substrate-aware network is the operating layer. Six organs — DePIN, DID, DeSci, DeFi, DPAi, DAOs — plus vCons as the nervous system. The anatomy that inhabits telecom, automotive, energy, and enterprise simultaneously. Designed for the cross-vertical pattern. Built for trust attestation. Operating at the substrate where AI-by-Design meets the listener.

Volume 3 coming.

Anchor on. 34°.

— Troy A. Resendez T.A. Resendez Incorporated

Sources (Consolidated)

- AI-RAN Alliance public documents and white papers (MWC 2024 forward)
- NVIDIA Aerial CUDA-Accelerated RAN documentation
- NVIDIA Drive Thor X technical documentation
- SoftBank AITRAS platform announcements (2024-2026)
- Mavenir AI-by-Design + MoVerge portfolio + Open RAN reflections from MWC 2026 / SATELLITE 2026
- NTT Data Managed Private 5G service documentation and case-study materials
- Tesla Foundation Model V0/V1 announcements and OpenXLA toolchain documentation
- Vision Language Action (VLA) model architecture references
- ITU-R IMT-2030 framework
- EU AI Act (Regulation 2024/1689)
- 3GPP Release 19 work items on AI/ML integration
- O-RAN Alliance Near-RT RIC and xApp/dApp framework specifications
- Distributed Energy Resources industry coverage (Wood Mackenzie, GTM Research)
- Virtual Power Plant operator case studies (California, Texas, New York deployments)
- Academic literature on SHAP, RuleFit, LoRA (Low-Rank Adaptation), and federated learning architectures
- vCons specification draft documents and IETF correspondence
- Industry coverage from RCR Wireless, Light Reading, FierceWireless

---

About The Substrate Report

The Substrate Report is a video podcast series from T.A. Resendez Incorporated, exploring the architectural convergence shaping the next layer of the internet. Each Volume is a three-episode arc covering one architectural thesis. Volumes are short-form (4-5 minutes per episode), substantively researched, and built for telecom-literate, technical, and Web3-adjacent audiences.

Volume 1 — The New Industrial Stack (June 2026): Spectrum war, AI at orbital speeds, DePIN counter-movement.

Volume 2 — AI by Design: The Cognitive RAN (July 2026): Design center shift, trust at the edge, cross-vertical pattern.

Volume 3 — Coming.

The hosts are Giga and Byte — the Substrate-Aware Network Correspondents — making their public debut in Volume 2 as the first pair in the planned Substrate-Aware Network Correspondents line. Their role is presence-architecture: providing AI + Human balance for the network's presence across physical and digital worlds.

The series is published at thesubstratereport.substack.com and on the @W3bHippo YouTube channel.

The Substrate Report · Volume 2 · AI by Design: The Cognitive RAN · A T.A. Resendez Inc. production · July 2026 · taresendezinc.com