InfraSense: A Sensing-Aware, Latency-Guaranteed AI Inference Layer for the 5G/6G RAN

AI-RAN Alliance Call for Innovation Proposals (2nd edition) — July 2026

---

1. Executive Summary

AI is moving into the radio access network faster than the network can safely host it. Inference workloads that drive real-time RAN control — beam management, interference mitigation, integrated sensing — carry hard latency bounds measured in milliseconds, yet today they are deployed onto whatever compute happens to be available, with no architectural guarantee that a response will arrive in time to be useful.

InfraSense is a working demonstrator of a *dedicated AI inference layer for the 5G/6G RAN* that makes latency a first-class, enforced property of every inference request, together with two companion innovations that make network *sensing* (ISAC) a manageable, intent-driven service:

1. **A dedicated inference layer**: every inference request declares a latency class; the orchestrator places it only on infrastructure tiers that can honor that class (DU-adjacent edge, RT RIC, Near-RT RIC, cloud) — and *refuses* a request rather than misroute it to a tier that cannot meet its bound.
2. **A1 "SensingIntent" policies**: a new A1 policy type that lets an operator express *what* the network should sense ("track this transmitter to 0.1 m at the stadium, without spending more than 50 % of radio resources on sensing") and have rApps/xApps derive and tune the radio configuration automatically in a closed loop.
3. **An SMO Global Agent**: a coordinator that detects when multiple intents and applications collectively over-commit the network, and resolves the conflict by priority — protecting critical sensing while trimming lower-priority consumers to declared floors.

All three components run today, end-to-end, in a live demonstrator on NVIDIA GPU infrastructure, with **measured (not simulated) evidence** for the core claim: in our tests, shared cloud inference latency varied by more than an order of magnitude run-to-run, while dedicated local GPU inference stayed within a narrow, predictable band across **two independent local serving engines** — precisely the variability a latency-guaranteed inference layer exists to master.

A US provisional patent application covering the architecture has been filed; the proposing team seeks AI-RAN Alliance collaboration to align the approach with Alliance working-group directions and mature it toward multi-vendor applicability.

2. Problem Statement and Market Relevance

The problem. Three converging trends collide in the 6G-bound RAN:

- **AI-for-RAN**: ML models increasingly make or inform RAN control decisions with strict deadlines (near-real-time loops of 10 ms – 1 s, real-time loops below 10 ms).

- **ISAC**: 3GPP and O-RAN roadmaps make the network itself a sensor — producing high-rate channel, reflection, and mobility data that must be processed close to where it is produced.
- **Heterogeneous compute**: usable inference capacity now spans DU-adjacent accelerators, RIC-hosted GPUs, and remote cloud — each with radically different latency behavior, especially under load.

Today there is no standard architectural component that binds these together. An xApp developer who needs a 10 ms inference answer has no portable way to *declare* that requirement and no guarantee the platform will honor it. The failure mode is silent: a model call that lands on a congested shared endpoint simply arrives late, and the control loop degrades unpredictably. Our measurements reproduce this failure mode on demand — public shared-cloud model endpoints in our tests swung from ~1 second to tens of seconds on identical requests, and during a measured congestion window (July 2026) individual shared-cloud calls stretched to 50–70 seconds with some requests rejected outright by the endpoint. Dedicated local GPU serving, by contrast, stayed within roughly 1.3–1.8× of its typical latency in every run — a behavior we have now reproduced on two independent local serving engines.

Equally, ISAC has no operator-grade management story: there is no standard way to tell the network *what* to sense, at what accuracy, at what resource cost — or to arbitrate when sensing goals collide with communication goals.

Market relevance. Every operator deploying AI-RAN capabilities faces this placement problem, and every ISAC use case (venue analytics, public safety, V2X perception, industrial digital twins) needs intent-level manageability before it can be sold as a service. The proposal addresses both gaps with mechanisms designed to sit on existing O-RAN interfaces (A1, E2, O1, O2) rather than replace them — a direct match to the Alliance's AI-for-RAN and AI-and-RAN work.

Positioning relative to the inaugural program. The Alliance's first innovation program showcased dynamic compute sharing between AI and RAN workloads on shared GPU hardware (e.g. GH200-based demonstrations). This proposal is deliberately positioned one layer above that work, and is made *more* necessary by it: the more aggressively RAN and AI workloads share accelerators, the less predictable any individual inference request's latency becomes — which is exactly the failure mode our measurements reproduce. InfraSense supplies the governance layer that dynamic compute sharing creates the need for: per-request latency classes, enforced tier placement, explicit refusal, and auditable control actions — plus the intent-driven sensing management dimension that no compute-sharing work addresses. It complements the inaugural program's innovations rather than competing with them.

3. Innovative Solution and Technical Approach

3.1 Dedicated inference layer (latency-guaranteed placement)

The inference layer receives requests comprising sensing data and a latency objective, classifies each request into a latency class, and selects an inference entity from a catalog — subject to a **hard tier rule**: a latency class admits only infrastructure tiers capable of honoring it (e.g. the strictest class admits only DU-adjacent / RT-RIC-adjacent entities). If no compliant placement exists, the request is **explicitly refused** rather than silently misrouted. Placement further considers

live infrastructure metrics, request priority (higher-priority requests can preempt modeled lower-priority load), and an accuracy objective: the orchestrator selects among model variants (fast / balanced / accurate) to meet accuracy within the latency bound, flagging when it cannot. Each run emits an auditable control action (beamforming, power, slice, handoff, sensing-schedule, or O-Cloud resource recommendation) over the interface appropriate to its kind and tier (E2 / O1 / O2).

3.2 A1 SensingIntent policies (intent-driven ISAC)

A new A1 policy type carries a sensing intent for a geographic/network scope: target metrics (e.g. localization accuracy, sensing coverage), resource constraints (e.g. a ceiling on radio resources or transmit power spent on sensing), a sensing mode, validity time, and priority. An rApp at the Non-RT RIC generates the policy object; the Near-RT RIC derives concrete sensing configuration from its parameters, applying the constraint caps; a closed loop then compares delivered sensing performance against the intent's targets and retunes. In the demonstrator this full chain — intent → A1 transmission → derived configuration with caps applied → closed-loop convergence toward the accuracy and coverage targets — runs live, using the worked examples from the filing (a 0.1 m critical tracking intent and a 90 % coverage intent).

3.3 SMO Global Agent (conflict arbitration)

A Global Agent at the SMO watches aggregate commitments across sensing intents and rApp/xApp resource consumers. When the sum over-commits the network (the demonstrator reproduces a 170 % commitment scenario), the agent resolves by priority: critical intents are protected in full, lower-priority intents are reduced no further than their declared floors, and non-sensing consumers are trimmed proportionally — with the resolution reported for audit. Where no complete resolution exists, infeasibility is reported explicitly rather than papered over.

3.4 Why this is innovative

The combination — *declared latency classes with refusal semantics*, *sensing as an A1-manageable intent*, and *priority arbitration at the SMO* — treats AI inference and sensing as first-class, governed network services rather than best-effort add-ons. The design philosophy is verifiable honesty: measured values are labelled measured, simulated values are labelled simulated, and the system fails explicitly instead of degrading silently. A US provisional application covering these mechanisms was filed in 2026; collaboration under this program would not be encumbered — the intent is to align the mechanisms with Alliance and O-RAN directions.

4. Deployment Feasibility

This is not a concept paper. All three components run today in an integrated demonstrator:

- **Live demonstrator**: a full-stack O-RAN-oriented dashboard (Node/React) in which the inference layer, SensingIntent policies, and Global Agent operate end-to-end — policies transmitted over A1, configurations derived with caps applied, the closed loop converging on targets, conflicts arbitrated by priority, and every control action auditable.
- **Real GPU execution**: inference runs on NVIDIA infrastructure — cloud NIM endpoints and dedicated local GPU serving. The latency evidence was measured as application-level round-trips of real model calls (not simulation) on two independent dedicated local serving stacks:

vLLM slices on A100 MIG hardware and, after a hardware refresh, a separate serving engine on an H100.

- **Measured core evidence**: a benchmark suite that has grown to cover every agent in the demonstrator's catalog (13 scenario tests as of July 2026), spanning both single-shot and genuine multi-step agentic model subjects. Repeated runs demonstrate the cloud-variability failure mode (order-of-magnitude swings and outright rejections on shared endpoints under congestion; narrow bands on dedicated local engines) that the placement rule exists to prevent. Every benchmark call is a real model call whose failures (timeouts, rejections, wrong answers) surface as explicit errors — nothing is retried into a pass — and runs are additionally traced on independent third-party infrastructure for verifiability.
- **Standards-aligned surfaces**: the mechanisms attach to A1, E2, O1, and O2 as defined by O-RAN — deployment means adding a policy type, an orchestrator, and an SMO service, not replacing the stack.

Remaining feasibility work is integration, not invention: aligning the SensingIntent information model with evolving O-RAN/3GPP ISAC work, exercising the layer against additional vendors' near-RT RICs, and hardening the demonstrator into a reference implementation.

5. 12-Month Timeline with Milestones

Months	Milestone	Exit criterion
M1–M2	**Baseline & methodology** — publish the demonstrator's benchmark methodology; freeze reproducible latency-evidence harness	Reproducible measured baseline report
M3–M5	**Alliance alignment** — map SensingIntent parameters and latency classes onto AI-RAN Alliance / O-RAN working-group terminology; incorporate feedback	Alignment note reviewed with Alliance members
M5–M8	**Interoperability** — exercise the inference layer and A1 policy flow against at least one additional (non-proprietary or partner) RIC implementation	Cross-implementation demo of intent → config → closed loop
M8–M10	**Field-shaped pilot** — run the sensing-intent chain in a realistic venue scenario (e.g. dense-venue localization + coverage intents with arbitration under load)	Pilot report with measured KPIs
M10–M12	**Reference release & report** — reference implementation documentation, final measured results, contribution summary to the Alliance	Final deliverable package

6. Expected Deliverables and Impact

Deliverables

1. Reproducible benchmark methodology + measured latency-evidence reports (shared-cloud vs dedicated-tier behavior).
2. A documented SensingIntent policy information model proposal aligned with Alliance/O-RAN vocabulary.
3. A working reference demonstrator of the full chain (intent → A1 → derived config → closed loop → arbitration → auditable control action).
4. Pilot report from the venue-shaped scenario.

5. Final report and working-group contribution summary.

Impact

- Gives the AI-RAN ecosystem a concrete, tested answer to "how do we *guarantee* AI inference latency in the RAN" — with refusal semantics that make failures visible instead of silent.
- Provides the missing operator-grade management story for ISAC: sensing as a priced, prioritized, intent-driven service — a prerequisite for every ISAC business case.
- Demonstrates a verifiable-honesty pattern (measured vs simulated always labelled) that raises the evidentiary bar for AI-RAN performance claims across the industry.

Contact: Haseem Ahmed, Co-Founder, DataSigns.AI — https://www.data-signs.ai/ · haseemahmed2@gmail.com · Richardson, TX, USA. *Patent notice: aspects of the described architecture are the subject of a filed US provisional patent application; the proposers are committed to collaboration under the Alliance's IPR framework.*