INFRASENSE

DATASIGNS.AI
CLOUD & ANALYTICS

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

THE PROBLEM

AI models increasingly drive real-time RAN control — but when an inference call lands on congested shared compute, it simply arrives late, and the control loop degrades silently. As AI-RAN moves compute sharing into the network, nothing today enforces the deadlines that RAN AI depends on.

MEASURED — NOT SIMULATED

Shared cloud endpoint

>10×

run-to-run latency swing on identical requests

Dedicated local GPU slice

1.4–1.8×

narrow, predictable band — application-level measurement

THE SOLUTION — THREE COMPONENTS

1 · Dedicated Inference Layer

Every request declares a latency class — placed only on tiers that can honor it, refused rather than misrouted. Every run emits an auditable control action.

2 · A1 SensingIntent Policies

Operators state what to sense, at what accuracy and resource cost — the network derives the configuration and a closed loop tunes it against the target.

3 · SMO Global Agent

Detects over-commitment across intents and applications — resolves by priority, protects critical sensing in full, and reports every resolution for audit.

ON THE OPEN RAN ARCHITECTURE

SMO / Non-RT RIC (hosts the Global Agent) → **A1 — SensingIntent policies** → **Near-RT RIC** (sensing xApps) → **E2 → CU / DU / RU**

The **Dedicated AI Inference Layer** sits beside the stack — cloud NIM endpoints and local GPU slices behind a latency-class gate — serving xApps and control loops.

170% aggregate over-commitment resolved live by the Global Agent — critical sensing protected in full, lower-priority consumers trimmed to declared floors, audit-reported.

✓ Working demonstrator ✓ Measured evidence ✓ Built for AI-RAN collaboration