AURORA: AI Unified RAN Orchestration & Resource Allocation with Agentic AI

Motivation

- **AI-RAN Convergence**

Future 6G networks will integrate AI across RAN control loops and infrastructure. AI-RAN enables three complementary directions: **AI-for-RAN** using AI to optimise the RAN, **AI-on-RAN** hosting AI services on the RAN, and **AI-and-RAN** enables AI and RAN workloads to share the same physical resources across edge, regional, and central clouds.

This shifts the RAN from managed communication infrastructure towards an intelligent, self-optimising edge compute fabric.

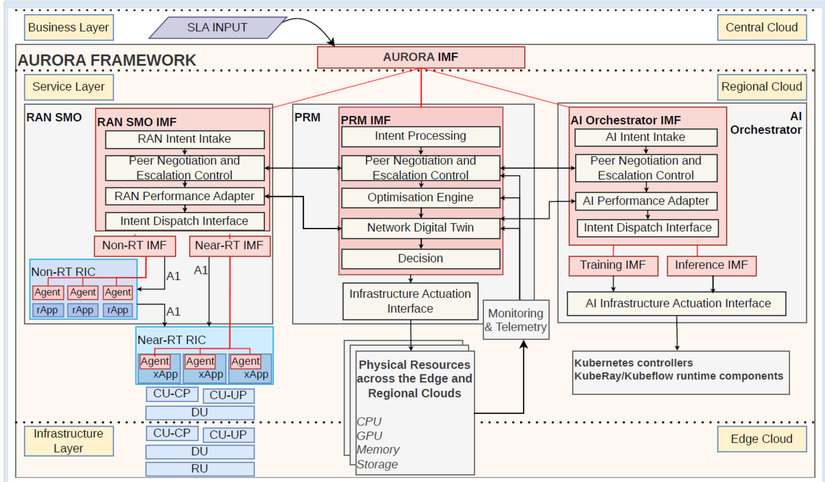
- **Need for Unified Agentic AI-RAN Orchestration**

Existing AI-native 6G architectures support intent-based management, service exposure, and NDT-assisted validation, but they do not provide a unified agentic AI-RAN orchestration framework. AI-RAN studies mainly address workload coexistence or resource allocation, while existing agentic AI approaches remain mostly limited to RAN-specific control loops. At the same time, heterogeneous interfaces, vendor extensions, and distributed control points create coordination conflicts at the RAN edge. AURORA addresses this gap through agentic AI coordination of AI workloads and RAN functions across edge, regional, and central clouds.

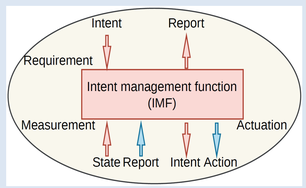
- **Design Questions and Contributions**

1. ***What*** workloads should be active, ***which*** policies should be enforced, and ***when*** should decisions be triggered?
   - ➢ Domain-specific IMFs for RAN and AI orchestration, coordinated through the AURORA IMF.
2. ***Where*** should workloads be executed, ***how*** should resources be allocated, and ***whether*** requested configurations are feasible?
   - ➢ PRM IMF as a unified placement and feasibility arbiter across the cloud continuum.
3. ***How*** should cross-domain conflicts be resolved when RAN and AI objectives compete?
   - ➢ AURORA IMF as the top-level arbitration layer for cross-domain conflict resolution.

AURORA Architecture



IMF-Based Orchestration

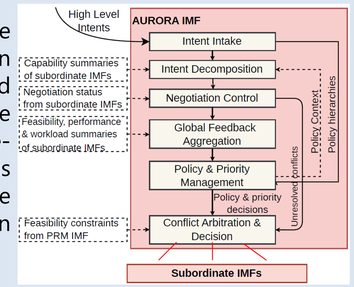


- An Intent Management Function (IMF) is an autonomous orchestration entity that observes network state, processes intents, makes decisions, and coordinates with peer IMFs within its assigned domain.
- In AURORA, the top-level AURORA IMF receives SLA inputs, decomposes high-level intents, and coordinates three subordinate IMFs: the RAN SMO IMF, PRM IMF, and AI Orchestrator IMF.
  - i. The RAN SMO IMF manages RAN intents and RIC-based control,
  - ii. The AI Orchestrator IMF manages AI workload and model lifecycle decisions,
  - iii. The PRM IMF acts as the shared resource arbiter across edge, regional, and central cloud resources.
- Together, these IMFs allow RAN and AI domains to operate independently while remaining coordinated through AURORA.

Agentic Coordination

- **Lease-Based Agent Actions**

AURORA uses agentic AI agents under the control of their parent IMFs. Any agent action that changes network state is first checked against constraints and conflict graphs. If the action is valid, the parent IMF grants a time-bounded lease; otherwise, the action is modified or rejected. This prevents race conditions and supports safe rollback when conflicts or failures occur.



- **Hierarchical Conflict Resolution**

Conflicts are first handled by subordinate IMFs through peer negotiation. If a conflict cannot be resolved locally, it is escalated to the AURORA IMF, which applies policy priorities and high-level business intents to make the final decision. This avoids circular negotiation and keeps AI and RAN decisions aligned with system-wide objectives.

- **Feasibility-Aware Resource Arbitration**

The PRM IMF validates workload placement and resource allocation before decisions are applied to the live network. Its optimisation engine generates candidate placements, while the Network Digital Twin checks their impact on RAN performance, AI service quality, and shared physical resources.

- **Validation Insight**

A 24-hour simulation validates the need for unified AI-RAN orchestration. Compared with single-domain RAN-Local and AI-Local baselines, AURORA reduces RAN and AI SLA violations while maintaining Edge-2 policy compliance.

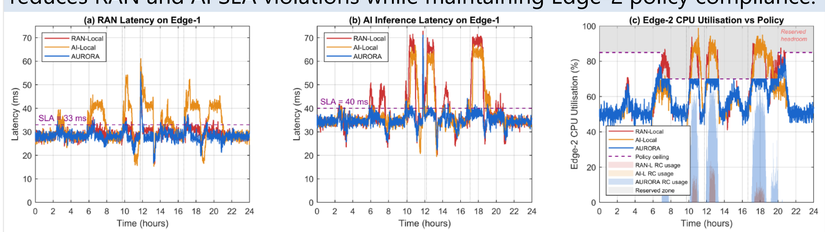


Fig. (a) RAN latency on Edge-1, (b) AI inference latency on Edge-1, and (c) Edge-2 CPU utilisation against the operator-defined reservation policy ceiling. Shaded areas in (c) indicate Regional Cloud usage per policy.

Alican Topcu*, Syed Ali Raza Zaidi*, Mallik Tatipamula† , Maryam Hafeez*

* School of Electronic and Electrical Engineering, University of Leeds, Leeds LS2 9JT, U.K.

† Ericsson, Santa Clara, California, USA

a.topcu@leeds.ac.uk