On Making AI-and-RAN Efficient and Safe

Presenters: Dr. Leyang Xue and Dr. Tianxin Wang
Advisor: Prof. Mahesh Marina
The University of Edinburgh

Future radio access networks (RANs) are moving toward AI-native compute infrastructure. As 6G architectures adopt server-grade accelerators for virtualized and programmable baseband processing, AI-RAN is emerging as a broader vision beyond applying AI to radio network optimization. Besides *AI-for-RAN*, aimed at improving energy, operational and spectral efficiencies, and *AI-on-RAN*, which is about hosting edge AI services on RAN infrastructure, *AI-and-RAN* asks whether the same RAN compute infrastructure can be shared safely between latency-critical RAN processing and non-RAN AI workloads. This direction is especially timely because GPU-accelerated RAN sites are geographically distributed and potentially in millions of scale, provisioned for peak traffic, and increasingly equipped with hardware capable of supporting modern AI computation.

We study this AI-and-RAN opportunity through the lens of foundation model (FM) training. FM training is a demanding workload: it requires significant GPU compute, mostly centralized in datacenters but in principle can benefit from opportunistic use of spare RAN compute capacity if that is exposed safely. However, RAN infrastructure cannot be treated as a conventional idle-compute pool. RAN processing is the first-class tenant: uplink decoding has sub-millisecond deadlines, and user performance cannot be traded for training throughput. The central question is therefore whether RAN compute provisioned-for-peak can be turned into a practical FM training substrate without compromising carrier-grade RAN behavior.

We first conduct a multi-scale characterization of GPU-accelerated RAN workloads. At the micro scale, RAN channel decoding is primarily memory-bandwidth-bound, while FM training is dominated by compute-bound matrix operations; the two workloads therefore stress different GPU resources and are more complementary than scalar utilization metrics may suggest. At the macro scale, measured cellular traffic shows that spare capacity is bursty within each site but complementary across sites, because traffic peaks are staggered across space and time. Across the characterized workloads and traffic conditions, roughly 40–85% of GPU compute remains unused, indicating that AI-RAN infrastructure can expose meaningful aggregate capacity if the sharing interface is fine grained, RAN safe and deadline aware.

Turning this spare capacity into a usable FM training substrate requires satisfying three requirements. First, RAN processing must be treated as the primary tenant, strictly respecting its function processing deadlines. Second, the exposed capacity must be stable enough for FM training kernels to exploit; directly following bursty per-slot RAN demand would force frequent GPU reconfiguration and waste much of the opportunity. Third, the training system must adapt across timescales, reacting quickly to slot-level capacity changes at one site while also redistributing work across sites as the available compute spatially shifts over longer time scales.

We address these challenges with our system, a RAN-first system for FM training over AI-RAN compute infrastructure. Rather than relying on an external resource manager to predict future radio demand, our system places spare-compute control inside the RAN stack. The key control point is the uplink MAC scheduler, whose resource-allocation decisions directly shape GPU decoding demand. By making this scheduler compute aware, our system exposes a stable and RAN-safe compute signal to co-located workloads while preserving per-slot decoding deadlines and user performance.

On top of this RAN-side interface, our system introduces a two-level elastic FM training runtime. Within each site, the runtime adapts training kernels to the local capacity signal, decomposing matrix operations into units that fit available GPU resources without blocking the next RAN slot. Across sites, a coordinator routes micro-batches toward sites with available capacity and changes the model layout only when persistent imbalance justifies the cost. A lightweight publish–subscribe interface couples the two layers: the RAN publishes safe spare-compute availability, and the training runtime continuously adapts to compute made available by the RAN after ensuring its deadlines are met.

We prototype this design on an O-RAN-aligned 5G NR stack with GPU-accelerated channel decoding and on a multi-site GPU-equipped cellular testbed. The RAN-side controller reduces spare-compute fluctuations by up to 4.9× while preserving HARQ deadline compliance and over-the-air user performance. When FM training consumes the stabilized envelope, the system harvests up to 83% of available spare compute and improves training throughput by 2.1–4.2× over coarse GPU sharing and dynamic training baselines. These results show that RAN compute infrastructure can become a practical decentralized AI training resource when spare capacity is exposed by the RAN itself and consumed by an elastic training system.