Zero-Lag Smart Pipes for Smart Factories: AI-Driven Programmable Transport in Open RAN

Flávio G C Rocha, Kleber V Cardoso, Murillo Melo, Jonathas dos Santos,
Lorenzo Chiachio, Vlademir Brusse, Fábio Luciano Verdi,
Leandro Almeida, Cristiano Bonato Both,
André Mendes Cavalcante, Maria Marquezini, Pedro Henrique Gomes
Federal University of São Carlos (UFSCar), Brazil
Universidade Federal de Goiás (UFG), Brazil
Federal Institute of Paraíba (IFPB), Brazil
University of Vale do Rio dos Sinos (Unisinos), Brazil
Ericsson Research, Brazil

***Abstract*—This demonstration addresses a key open challenge in Open Radio Access Network (O-RAN) deployments: how to intelligently allocate Transport Network (TN) resources to ensure low-latency for mission-critical applications. The demo emulates a Smart Factory scenario where the time-sensitive control traffic of robotic arms competes with industrial camera broadband video streams. We propose an intelligent transport controller that combines network slicing, Adaptive Neuro-Fuzzy Inference System (ANFIS), and Federated Learning (FL) to dynamically prioritize traffic per slice. The architecture uses P4 switches for local queue monitoring and real-time resource scheduling. The integration with the O-RAN disaggregated stack is based on Open Air Interface (OAI). Experimental results demonstrate valuable load balancing and buffer occupation reduction in the O-RAN midhaul.**

I. Introduction

As industries evolve towards more interconnected and automated systems under the Industry 4.0 paradigm, private 5G networks are increasingly being adopted [1]. One of the emerging deployment strategies for these networks is the Open Radio Access Network (O-RAN) architecture defined by the O-RAN Alliance [2]. This architecture is highly disaggregated, with RAN components distributed across geographical locations and *cloudification* becoming the preferred approach [3]. In this context, a critical open question arises: *how can programmable Transport Networks (TNs) and AI-driven automation be leveraged to achieve low-latency communication for industrial services across disaggregated O-RAN deployments?*

This challenging research question becomes even more pressing due to the dense and heterogeneous nature of converged media access, where fixed and mobile communication technologies share the same infrastructure, even in the last mile [4]. In such scenarios, mission-critical services will require prioritization through more intelligent and adaptive mechanisms, as traditional static configurations will no longer be sufficient to meet these emerging demands [5]. To address this, the network must adopt intelligent and distributed resource allocation strategies to ensure strict Quality of Service (QoS) guarantees, particularly under Ultra-Reliable and Low-Latency Communications (URLLC) requirements.

While O-RAN provides the RAN Intelligent Controller (RIC) to introduce intelligence into the RAN domain, these agents do not interact with TN nodes. Therefore, a dedicated controller for the transport segment capable of having a global view of the network and acting in a distributed manner is essential to achieve end-to-end service provisioning in industrial scenarios with disaggregated RANs. This demonstration shows that such integration is feasible by combining the network slicing paradigm, the TN programmability relying on Software Defined Network (SDN) principles, the Adaptive Neuro-Fuzzy Inference System (ANFIS), and a distributed intelligence approach using Federated Learning (FL), forming a cohesive architecture for intelligent and adaptive resource allocation.

II. System Architecture

The architecture consists of two domains: TN and RAN, as depicted in Fig. 1. The TN domain comprises Tofino P4 programmable switches, acting as intelligent agents [6] and collecting real-time queue metrics as input to the ANFIS algorithm. The ANFIS agent deployed at each of these nodes dynamically adjusts the parameters of the Gaussian membership functions, a core component of Fuzzy logic. Each agent deployed on a Tofino switch undergoes a learning process based on the backpropagation algorithm, optimizing the mean and standard deviation of each membership function to achieve improved resource scheduling performance. However, having only a local view on each switch is insufficient for a global joint observation of the entire network. Our proposal leverages an FL algorithm to obtain a macro view of the TN. The local decision is reported to a centralized FL node, which aggregates and processes updates for each agent in a distributed learning manner. Different from other distributed learning approaches, the FL was adopted to reduce overhead in transmissions from local nodes to the centralized FL unit.

The FL ensures that only model parameters are exchanged, maintaining privacy and efficiency, key characteristics for open architectures, where multi-vendor devices are expected

978-3-903176-75-1 ©2025 IFIP

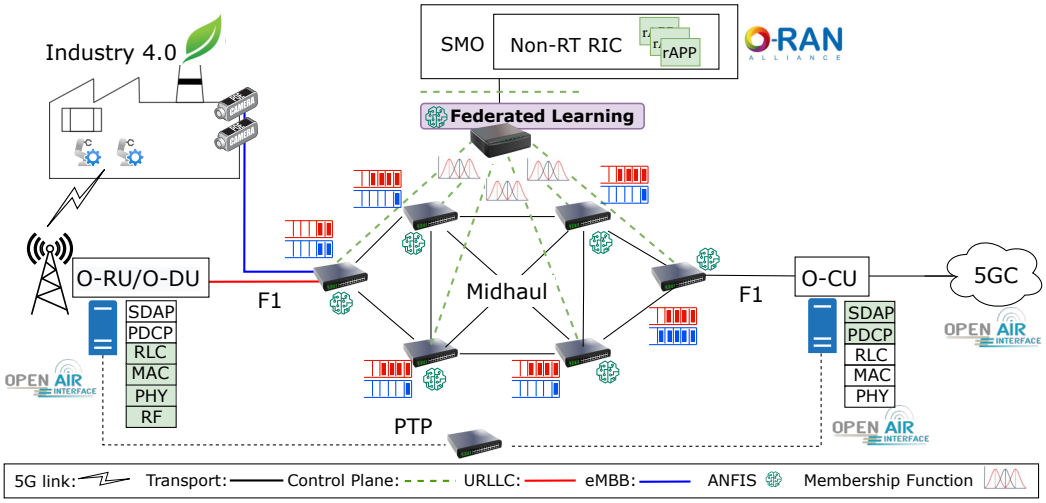

Fig. 1. Open RAN Transport Architecture.

to coexist. Moreover, our proposal was designed to remain operational in the event of a centralized controller failure, relying on local learning mechanisms at each node. While the FL controller enhances the system's performance by providing a macro-level view and global observability of the network, the system is capable of gracefully reducing to a decentralized mode, where nodes continue to operate autonomously. During such failures, the system still maintains reasonable performance, ensuring load balancing across network slices per node until the centralized controller is restored. This fault-tolerance capability represents a key strength of our approach, as it combines the benefits of global coordination with the robustness of localized decision-making based solely on local agent operation.

Algorithm 1. This algorithm describes the operation of our Federated ANFIS-based resource allocation framework, designed to perform intelligent and adaptive traffic scheduling in the TN, while exhibiting robustness against the failure of the centralized controller. Each programmable Tofino switch hosts a local ANFIS agent that continuously monitors its queue occupancy for each network slice. Based on this local state, the ANFIS agent computes updated Weighted Round Robin (WRR) weights, dynamically adapting the share of service allocated to each slice.

The learning process involves tuning the parameters of the fuzzy membership functions, specifically the means (μ) and standard deviations (σ) of Gaussian membership functions, via gradient-based optimization. These updates reflect changes in the traffic conditions observed locally by each switch. When the centralized FL controller is available, each switch sends its locally modeled parameters (μ and σ) to the controller. The FL controller performs model parameter aggregation using the FedAVG algorithm, i.e., via weighted averaging, generating a refined global model that captures network-wide patterns. This aggregated model is sent back to the switches to update their local ANFIS agents. If the FL controller becomes unavailable, each switch autonomously continues to operate using only its locally trained ANFIS model. This fallback mode ensures uninterrupted operation of the network, preserving acceptable load balancing and QoS enforcement without global coordination. Once the controller is restored, the system resumes normal federated operation by synchronizing local and global models.

Algorithm 1 Federated ANFIS-based Resource Allocation

Require: Transport network with N Tofino switches, each with a local ANFIS agent, and Central FL Controller

repeat
  for each switch $i \in \{1, \ldots, N\}$ **in parallel do**
    Monitor queue occupancies q_i^s for each slice s
    ANFIS computes local WRR weights ϕ_i^s based on q_i^s
    Update Gaussian membership parameters μ_i^l, σ_i^l via backpropagation for all Fuzzy linguistic levels l
    if FL controller is available **then**
      Send local ANFIS parameters to FL controller
    end if
  end for
  if FL controller is available **then**
    FL controller aggregates all models and updates global ANFIS parameters
    for each switch $i \in \{1, \ldots, N\}$ **do**
      Receive updated parameters $\mu_i^{l'}$, $\sigma_i^{l'}$ from FL
      Update local ANFIS model
      Apply updated WRR weights $\phi_i^{s'}$ for each slice s
    end for
  else
    Switches operate in fallback mode, utilizing local ANFIS agents, and leveraging local data only for resource allocation decisions
  end if
until End of operation

III. Testbed & Workflow

The demonstration environment is built upon a small-scale yet representative TN topology, designed to emulate realistic conditions in an O-RAN-based system. This topology consists of programmable Intel Tofino switches and three high-performance rack servers, each serving a specific role in the deployment. On the first rack server, we have instantiated the O-RAN Radio Unit (O-RU) and the O-RAN Distributed Unit

(O-DU), using the open-source 5G radio stack provided by the OAI project [7]. This server acts as the first point of radio signal emulation and processing, leveraging the `RFsimulation` feature of the OAI. The network traffic transmitted through this server represents a network slice in our scenario, corresponding to multiple URLLC flows, associated with a time-critical application: motion control of smart factory robots. We generated this traffic programmatically using a Python script to launch multiple parallel `iperf3` instances, simulating simultaneous URLLC connections to the 5G Core (5GC). The generated packets are injected into the network stack via an emulated User Equipment (UE) instance, under OAI.

The second rack server operates as the source of video traffic, mimicking high-bandwidth industrial cameras. This traffic generator represents another network slice considered in our scenario, dedicated to enhanced Mobile BroadBand (eMBB) services. This server hosts an `Apache` web server that delivers a high-resolution video stream using `MPEG-DASH`. On the third rack server, a JavaScript-based client application accesses and plays this stream, with the traffic being routed into the TN via the first hop of the Tofino-based switching fabric. To create a heterogeneous traffic scenario, this video traffic bypasses the 5G protocol stack and is injected directly into the first hop of the TN. It highlights the complexity foreseen for the disaggregated Open RAN interfaces, where traffic from different sources compete for shared network resources.

The third rack server hosts the remaining 5G network components, including the O-RAN Central Unit (O-CU) and the 5GC, which are deployed using the OAI framework. This server receives and processes control and user plane data, completing the end-to-end communication chain. It is responsible for consuming the incoming video stream via a media player application, enabling real-time monitoring and quality assessment.

To interconnect the three servers, we deployed the O-RAN High Layer Split (HLS), disaggregating the O-DU from the O-CU and linking them through a multi-node topology composed of Intel Tofino switches. This setup emulates the O-RAN F1 interface, comprising the F1-C (control plane) and F1-U (user plane) segments. Specifically, the data and signaling exchange between the first two servers and the third one is carried out via these F1 interfaces, accurately representing a real-world disaggregated RAN deployment where the O-DU and O-CU are deployed on separate physical nodes.

Figure 1 illustrates the described setup, showing the logical and physical interconnection of all components and highlighting how traffic flows from distinct network slices are mapped for different queues at each transport node and transmitted through the entire TN. This environment enables us to assess the behavior of the network under heterogeneous traffic conditions, evaluate QoS parameters, and validate the effectiveness of the proposed slice-aware and ANFIS-based transport resource allocation.

During the demo. Attendees will experience firsthand the potential benefits of employing intelligent resource allocation in the O-RAN TN. The step-by-step logic and parameterization of **Algorithm 1** will be illustrated during the demo workflow. A dashboard (see Fig. 2) will present the network topology at the top and, from left to right at the bottom, the queue occupancies, the WRR weights computed by the Fed-ANFIS algorithm, and an animated view of the queue states. The observed results include reduced latency for URLLC traffic, more efficient utilization of midhaul capacity, as well as the framework's resilience to events of centralized FL controller failure.

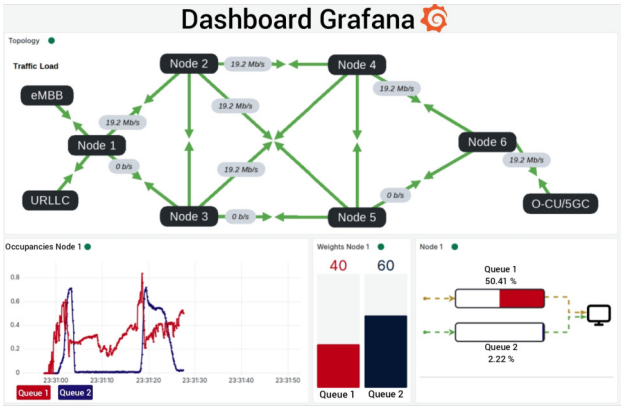

Fig. 2. Dashboard displaying the network topology, real-time buffer occupancies, and dynamic WRR decisions.

IV. Conclusions

The demo outcomes validate the feasibility of applying fuzzy-based distributed intelligence in real-time industrial networks. Moreover, the demo demonstrates how an open-source OAI deployment combined with P4-programmable switches can effectively emulate the targeted industrial communication environment, enabling valuable analysis in the context of disaggregated 5G networks. In addition, the demo can also be helpful for future 6G developments involving heterogeneous access networks.

Acknowledgements

This work was supported by Ericsson Telecomunicações Ltda., and by the São Paulo Research Foundation (FAPESP), grant 2021/00199-8, CPE SMARTNESS.

References

[1] S. M. Darroudi *et al.*, "On the TSN and 5G network integration approaches, 5G features proof, advantages and challenges," in *2024 Joint European Conference on Networks and Communications & 6G Summit (EuCNC/6G Summit)*. IEEE, 2024, pp. 688–693.

[2] O-RAN ALLIANCE, "O-RAN Specifications," https://specifications.o-ran.org/specifications, 2025, accessed: 2025-05-17.

[3] M. A. Habibi *et al.*, "Towards a fully virtualized, cloudified, and slicing-aware RAN for 6G mobile networks," in *6G Mobile Wireless Networks*. Springer, 2021, pp. 327–358.

[4] J. Khan and L. Jacob, "Resource allocation for CoMP enabled URLLC in 5G C-RAN architecture," *IEEE Systems Journal*, vol. 15, no. 4, pp. 4864–4875, 2020.

[5] W. Guan *et al.*, "RAN Slicing Towards Coexistence of Time-Sensitive Networking and Wireless Networking," *IEEE Communications Magazine*, 2023.

[6] D. Scano *et al.*, "Extending P4 in-Band Telemetry to User Equipment for Latency-and Localization-aware Autonomous Networking with AI Forecasting," *Journal of Optical Communications and Networking*, vol. 13, no. 9, pp. D103–D114, 2021.

[7] OpenAirInterface Software Alliance, "OpenAirInterface," https://openairinterface.org/, 2025, accessed: 2025-05-17.