CloudRIC demo: Open Radio Access Network (O-RAN) Virtualization with Shared Heterogeneous Computing

Leonardo Lo Schiavo*§, Gines Garcia-Aviles†, Andres Garcia-Saavedra‡
Marco Gramaglia§, Marco Fiore*, Albert Banchs*§, Xavier Costa-Perez†‡¶
*IMDEA Networks Institute, †i2CAT, ‡NEC Laboratories Europe, §Universidad Carlos III de Madrid, ¶ICREA

ABSTRACT

Open and virtualized Radio Access Networks (vRANs) are breeding a new market with unprecedented opportunities. However, carrier-grade vRANs today are expensive and energy-hungry, as they rely on hardware accelerators (HAs) that are dedicated to individual distributed units (DUs). We demonstrate CloudRIC [17], a system that, powered by lightweight data-driven models, meets specific reliability targets while (*i*) coordinating access between DUs and heterogeneous computing infrastructure; and (*ii*) assisting DUs with compute-aware radio scheduling procedures. Using a user-friendly dashboard to control an experimental testbed remotely, we demonstrate that CloudRIC achieves comparable reliability performance to a DU-dedicated platform while offering up to 40x higher cost-efficiency and up to 6x higher energy efficiency when pooling resources for up to 70 DUs.

CCS CONCEPTS

• **Networks** → **Mobile networks**; **Network reliability**.

KEYWORDS

vRAN, O-RAN, O-Cloud, Distributed Unit, HW Accelerators

ACM Reference Format:
Leonardo Lo Schiavo*§, Gines Garcia-Aviles†, Andres Garcia-Saavedra‡, Marco Gramaglia§, Marco Fiore*, Albert Banchs*§, Xavier Costa-Perez†‡¶. 2024. CloudRIC demo: Open Radio Access Network (O-RAN) Virtualization with Shared Heterogeneous Computing. In *The 30th International Conference on Mobile Computing and Networking (ACM MobiCom '24), November 18-22, 2024, Washington D.C., USA.* ACM, New York, NY, USA, 3 pages. https://doi.org/10.1145/3636534.3698858

1 INTRODUCTION

The rapid growth of open and virtualized Radio Access Networks (vRANs) presents a significant opportunity in the telecommunications industry. The flexibility of vRANs, which allows baseband processing on standard servers, offers benefits such as reduced vendor lock-in and simplified upgrades. However, the current implementation of vRANs faces challenges in terms of cost and energy efficiency.

The demanding computational requirements of 5G, particularly in tasks like forward error correction (FEC), need dedicated hardware accelerators (HAs) to meet strict latency and reliability targets [10, 14]. These HAs, while effective in improving performance, are expensive and energy-hungry. The prevailing industry practice of dedicating HAs to individual distributed units (DUs) further exacerbates these cost and energy concerns.

CloudRIC [17] enhances the efficiency of vRANs by sharing pools of heterogeneous processors among DUs. The approach leverages the observation that CPUs, while not capable of handling all 5G workloads alone, can effectively complement HAs for some of them. By opportunistically offloading less demanding loads to CPUs and sharing HAs among DUs, significant cost and energy savings can be achieved.

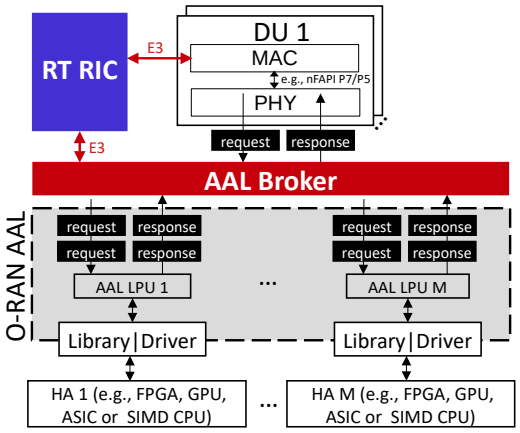

Figure 1: CloudRIC-powered O-Cloud high-level architecture. CloudRIC extensions are highlighted in red and blue.

2 CLOUDRIC

To this end, CloudRIC introduces a brokering system that coordinates access to shared resources and optimizes radio scheduling policies in real-time. CloudRIC seamlessly integrates into the O-RAN architecture and employs lightweight

ACM MobiCom '24, November 18-22, 2024, Washington D.C., USA

data-driven models to ensure reliability and efficiency. The core of CloudRIC's design lies in two key enhancements to the standard O-RAN architecture:

- Real-Time RIC (RT-RIC): Using a light actor-critic learning agent, the RT-RIC audits radio grants issued by DUs to ensure timely processing of all scheduled Transport Blocks (TBs) by the O-Cloud. It assists DUs with compute-aware radio policies in real-time, addressing the limitations of existing O-RAN RICs, and can be implemented with frameworks such as [8, 13].
- AAL Broker (AAL-B): The AAL-B provides DUs with a unified abstraction of the O-Cloud, enabling centralized coordination of its processors. It consists of two sub-components:
  - AAL-B User Plane (AAL-B-UP): Acts as a proxy between O-RAN Network Functions (NFs) like DUs and the O-RAN Acceleration Abstraction Layer (AAL), routing TBs to appropriate Logical Processing Units (LPUs) assigned by the AAL-B Control Plane.
  - AAL-B Control Plane (AAL-B-CP): Uses light data-driven LPU models to schedule granted TBs to LPUs, ensuring PHY processing deadlines are met at the minimum energy cost.

CloudRIC's design facilitates efficient load balancing across heterogeneous processors, ensuring reliability by proactively adapting workload to processing capacity through compute-aware radio policies. The decoupling of AAL-B-CP and AALB-UP minimizes data-plane overhead, further enhancing the system's efficiency. The detailed architecture and workflow of CloudRIC can be found in [17].

3 DEMONSTRATION

Fig. 2 illustrates our implementation of CloudRIC in an Intel server with two LPUs: an LPU based on an NVIDIA V100 GPU, and an LPU based on a pool of 16 Intel Xeon Gold CPU cores. The system involves the development of both the user plane (AAL-B-UP) and the control plane (AAL-B-CP and RT-RIC). The AAL-B-UP builds on DPDK's EAL, efficiently routes TBs to their assigned LPU queues. The CPU and GPU LPUs operate as separate threads, handling tasks like fetching, enqueuing, processing, and outputting decoded data. The control plane, implemented in C++, utilizes DPDK for inter-process communication and ONNX Runtime for neural network acceleration. The AAL-B-CP manages the scheduling pipeline, queue updates, time estimation, and LPU ML models. The RT-RIC employs a light actor-critic learning agent for radio policy decisions. LPU models, realized as feed-forward neural networks, are trained offline with a custom loss function to prioritize accurate latency estimation.

To demonstrate CloudRIC's capabilities, we compare its performance against five benchmarks:

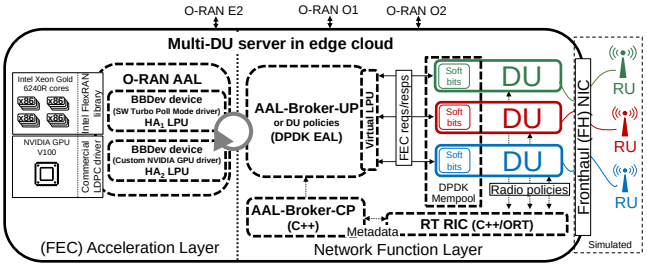

Figure 2: A CloudRIC-powered O-Cloud prototype.

Figure 3: CloudRIC demo dashboard.

- "Industry-std": Simulates a dedicated server and hardware accelerator (HA) per DU to guarantee reliability.
- "O-Greedy": DUs greedily select the HA first for every TB.
- "O-MWT": DUs use a data-driven LPU model to select the LPU that minimizes waiting time for processing each TB.
- "C-MWT": Extends O-MWT by centrally executing MWT using the AAL-Broker and information from all DUs.
- "C-DA": Extends C-MWT, with the AAL-broker selecting the lowest-energy-consuming LPU predicted to meet each TB's deadline.

We evaluate these approaches based on reliability (share of TBs timely processed), network throughput (Mb/s), energy efficiency (Mb/milli-joule), and cost-efficiency (Mbps/$).

The demonstration is controlled via a user-friendly dashboard (Fig. 3), which remotely manages the system. It allows for deploying a variable number of DU instances with different workload types (based on traces from real base stations, as explained in [17]) and selecting the solution to be employed (CloudRIC or a benchmark). Additionally, the dashboard collects and presents performance metrics in real-time in a user-friendly manner.

4 RELATED WORK

The use of CPU-based software processors for 5G workloads have raised substantial interest recently. Agora [4], a 5G data channel processor leveraging FlexRAN's libraries, operates on multi-core CPUs without necessitating specialized

HAs. However, its reliance on dedicated cores for deterministic latency hinders its applicability in shared computing scenarios. Concordia [7] introduces a CPU scheduler for concurrent 5G task and latency-elastic application execution. Nevertheless, it falls short in (*i*) enforcing radio resource control for multi-DU reliability and (*ii*) handling diverse HAs and CPUs. Nuberu [11], a DU prioritizing reliability over network latency in shared platforms, lacks mechanisms for coordinating multiple DUs or optimizing resource utilization. Further, GPF [12], a rapid GPU-based radio scheduler, and the approach in [15] for RU front-end sharing, do not explicitly address the combined control of computing and radio resources, which is central to our work.

Driven by O-RAN's openness, a number of research efforts have studied the use of machine learning for medium to long-term optimization (e.g., [1–3]), and platforms like ColO-RAN [16] and OrchestRAN [5] facilitate intent-based control and algorithm evaluation. However, their operational timescales (seconds or minutes) are misaligned with the short-term burstiness observed in real-world workloads (see [17]). Consequently, they cannot achieve the multiplexing gains of CloudRIC while upholding strict reliability guarantees.

In contrast to these longer timescales, the concept of *dApps* was proposed in [6] as micro-services for real-time inference for RAN applications, similar to our RT-RIC. Janus [9] proposes a series of hooks and codelets within each DU for real-time intelligence. Finally, EdgeRIC [13] adopts an alternative approach that is decoupled from the RAN stack, using O-RAN compliant messaging for vendor-agnostic real-time RAN control.

ACKNOWLEDGMENTS

Work supported by the EC via grants no. 101139270 (ORIGAMI) and by NextGeneration EU through UNICO I+D grants no. TSI-063000-2021 (OPEN6G) and 022/0005395 (CLARION).

REFERENCES

[1] Jose A. Ayala-Romero et al. 2020. vrAIn: Deep Learning based Orchestration for Computing and Radio Resources in vRANs. *IEEE Transactions on Mobile Computing* (2020), 1–1. https://doi.org/10.1109/TMC.2020.3043100

[2] Jose A. Ayala-Romero et al. 2021. Bayesian Online Learning for Energy-Aware Resource Orchestration in Virtualized RANs. In *IEEE INFOCOM 2021 - IEEE Conference on Computer Communications*. 1–10. https://doi.org/10.1109/INFOCOM42981.2021.9488845

[3] Jose A. Ayala-Romero et al. 2021. *EdgeBOL: Automating Energy-Savings for Mobile Edge AI.* Association for Computing Machinery, New York, NY, USA, 397–410. https://doi.org/10.1145/3485983.3494849

[4] Jian Ding, Rahman Doost-Mohammady, Anuj Kalia, and Lin Zhong. 2020. Agora: Real-time massive MIMO baseband processing in software. In *Proceedings of ACM CoNEXT '20*. ACM.

[5] Salvatore D'Oro, Leonardo Bonati, Michele Polese, and Tommaso Melodia. 2022. OrchestRAN: Network Automation through Orchestrated Intelligence in the Open RAN. In *IEEE INFOCOM 2022 - IEEE Conference on Computer Communications*. 270–279. https://doi.org/10.1109/INFOCOM48880.2022.9796744

[6] Salvatore D'Oro, Michele Polese, Leonardo Bonati, Hai Cheng, and Tommaso Melodia. 2022. dApps: Distributed Applications for Real-Time Inference and Control in O-RAN. *IEEE Communications Magazine* 60, 11 (2022), 52–58. https://doi.org/10.1109/MCOM.002.2200079

[7] Xenofon Foukas and Bozidar Radunovic. 2021. Concordia: Teaching the 5G VRAN to Share Compute. In *Proceedings of the 2021 ACM SIGCOMM 2021 Conference* (Virtual Event, USA) *(SIGCOMM '21)*. Association for Computing Machinery, New York, NY, USA, 580–596. https://doi.org/10.1145/3452296.3472894

[8] Xenofon Foukas, Bozidar Radunovic, Matthew Balkwill, and Zhihua Lai. 2023. Taking 5G RAN Analytics and Control to a New Level. In *Proceedings of the 29th Annual International Conference on Mobile Computing and Networking* (Madrid, Spain) *(ACM MobiCom '23)*. Association for Computing Machinery, New York, NY, USA, Article 1, 16 pages. https://doi.org/10.1145/3570361.3592493

[9] Xenofon Foukas, Bozidar Radunovic, Matthew Balkwill, Zhihua Lai, and Connor Settle. 2023. Programmable RAN Platform for Flexible Real-Time Control and Telemetry. In *Proceedings of the 29th Annual International Conference on Mobile Computing and Networking*. 1–3.

[10] Fujitsu. 2023. What is the difference between inline and lookaside accelerators in virtualized distributed units? *White Paper* (2023).

[11] Gines Garcia-Aviles, Andres Garcia-Saavedra, Marco Gramaglia, Xavier Costa-Perez, Pablo Serrano, and Albert Banchs. 2021. Nuberu: Reliable RAN Virtualization in Shared Platforms. In *Proceedings of the 27th Annual International Conference on Mobile Computing and Networking* (New Orleans, Louisiana) *(MobiCom '21)*. Association for Computing Machinery, New York, NY, USA, 749–761. https://doi.org/10.1145/3447993.3483266

[12] Yan Huang, Shaoran Li, Y. Thomas Hou, and Wenjing Lou. 2018. GPF: A GPU-Based Design to Achieve 100 μs Scheduling for 5G NR. In *Proceedings of the 24th Annual International Conference on Mobile Computing and Networking* (New Delhi, India) *(MobiCom '18)*. Association for Computing Machinery, New York, NY, USA, 207–222. https://doi.org/10.1145/3241539.3241552

[13] Woo-Hyun Ko, Ushasi Ghosh, Ujwal Dinesha, Raini Wu, Srinivas Shakkottai, and Dinesh Bharadia. 2024. EdgeRIC: Empowering Real-time Intelligent Optimization and Control in NextG Cellular Networks. In *21st USENIX Symposium on Networked Systems Design and Implementation (NSDI 24)*. USENIX Association, Santa Clara, CA, 1315–1330.

[14] Light Reading. 2023. *Chip choices kickstart open RAN war between lookaside and inline.* https://www.lightreading.com/semiconductors/chip-choices-kickstart-open-ran-war-between-lookaside-and-inline

[15] Jose Mendes, XianJun Jiao, Andres Garcia-Saavedra, Felipe Huici, and Ingrid Moerman. 2019. Cellular access multi-tenancy through small-cell virtualization and common RF front-end sharing. *Computer Communications* 133 (2019), 59–66. https://doi.org/10.1016/j.comcom.2018.10.010

[16] Michele Polese, Leonardo Bonati, Salvatore D'Oro, Stefano Basagni, and Tommaso Melodia. 2022. ColO-RAN: Developing Machine Learning-based xApps for Open RAN Closed-loop Control on Programmable Experimental Platforms. *IEEE Transactions on Mobile Computing* (2022), 1–14. https://doi.org/10.1109/TMC.2022.3188013

[17] Leonardo Lo Schiavo, Gines Garcia-Aviles, Andres Garcia-Saavedra, Marco Gramaglia, Marco Fiore, Albert Banchs, and Xavier Costa-Perez. 2024. CloudRIC: Open Radio Access Network (O-RAN) Virtualization with Shared Heterogeneous Computing. In *Proceedings of the 30th Annual International Conference on Mobile Computing and Networking*. 558–572. https://doi.org/10.1145/3636534.3649381