Financiado por la Unión Europea
NextGenerationEU

MINISTERIO
DE ASUNTOS ECONÓMICOS
Y TRANSFORMACIÓN DIGITAL

Plan de Recuperación, Transformación y Resiliencia

uc3m

6G-CLARION-NFD Entregable E8

Use cases definition

PROGRAMA DE UNIVERSALIZACIÓN DE INFRAESTRUCTURAS DIGITALES PARA LA COHESIÓN

UNICO I+D 5G 2021

UNICO I+D

Fecha: 31/7/2022

Versión: 1.0

Propiedades del documento

Id del documento	E8
Título	Use cases definition
Responsable	UC3M
Editor	Albert Banchs

Equipo editorial	**Partner**	**Name**	**Surname**	**Sections**
	UC3M	Albert	Banchs	All

Nivel de diseminación	Público
Estado del documento	Final
Versión	1.0

Historial

Revisión	**Fecha**	**Por**	**Descripción**
1.0	31/07/22	Editor	Final version

Revisor

Equipo revisor	**Partner**	**Name**	**Surname**	**Sections**
	UC3M	Marco	Gramaglia	All

Descargo de responsabilidad

This document has been produced in the context of the 6G-CLARION Project. The research leading to these results has received funding from the Spanish Ministry of Economic Affairs and Digital Transformation and the European Union-NextGenerationEU through the UNICO 5G I+D programme.

Table of Contents

Lista de acrónimos

BSR: Buffer State Report

CRC: Cyclic Redundancy Checks

DL: Downlink

gNB: Next-Generation Node B

HARQ: Hybrid Automatic Repeat Request

MCS: Modulation and Coding Scheme

PRB: Physical Resource Block

SNR: Signal-to-Noise Ratio

TBS: Transport Block Size

TTI: Transmission Time Interval

UL: Uplink

UE: User Equipment

vRAN: Virtualized Radio Access Network

Resumen ejecutivo

La virtualización de RAN se convertirá en una tecnología clave para el último tramo de las redes móviles de próxima generación impulsadas por iniciativas como la alianza O-RAN. Sin embargo, debido a las fluctuaciones informáticas inherentes a la dinámica inalámbrica y la contención de recursos en la infraestructura informática compartida, el precio de migrar de plataformas dedicadas a compartidas puede ser demasiado alto. De hecho, mostramos en este documento que la arquitectura de línea de base de la unidad distribuida (DU) de una estación base puede sufrir momentos de déficit en capacidad de cómputo. Por lo tanto, se debe diseñar una nueva arquitectura de canalización para las DUs 4G/5G específicamente diseñada para plataformas informáticas no deterministas. En este documento hacemos el caso para tal aplicación..

Abstract

RAN virtualization is expected to play a crucial role in the final phase of the next-generation mobile networks, with initiatives like the O-RAN alliance driving the adoption of this technology. However, the transition from dedicated to shared platforms may pose significant challenges due to the inherent computing fluctuations in wireless dynamics and resource contention in shared computing infrastructure. This can result in the price of migration being too high for some network operators. This document highlights the potential risks and deficiencies in the baseline architecture of a base station's distributed unit (DU), which can suffer from moments of deficit in computing capacity. To address these issues, we propose that a novel pipeline architecture specifically engineered for non-deterministic computing platforms, should improve the performance and reliability of 4G/5G DUs. Our proposed architecture can help to mitigate the risks associated with RAN virtualization and enable network operators to achieve optimal performance in their networks. In this document, we present a strong case for the adoption of our proposed pipeline architecture for 4G/5G DUs on non-deterministic computing platforms.

1. Introduction

The virtualization of radio access networks (RANs) is rapidly evolving and is expected to be a key aspect of next-generation mobile systems beyond 5G. Currently, RANs rely on monolithic appliances over ASICs. However, initiatives such as the O-RAN alliance have stimulated the market and research community to develop innovative solutions that incorporate the flexibility and cost-efficiency of network function virtualization (NFV) into the edge of mobile networks.

A virtualized RAN (vRAN)[1] architecture includes base stations (BSs) that are split into a central unit (CU), distributed unit (DU), and radio unit (RU). The highest layers of the stack are hosted in the CU, while the physical layer (PHY) is hosted in the DU, and basic radio functions such as amplification or sampling are hosted in the RU. The vRAN architecture relies on cloud platforms that have shared computing resources to host virtualized functions such as the PHY.

However, while CUs can be virtualized in regional clouds, virtualized DUs (vDUs) require fast and predictable computation in edge clouds, particularly the vPHY. Shared computing platforms do not provide the necessary predictability for DUs, which trade off the predictability provided by dedicated platforms for higher flexibility and cost-efficiency.

Research has shown that resource contention in shared computing infrastructure may lead to up to 40% performance degradation compared to dedicated platforms, a phenomenon known as the "noisy neighbor" problem. This issue is particularly relevant for traditional network functions such as virtual switches, firewalls, or even CUs, where metrics such as tail latency are crucial. Therefore, substantial effort has been devoted to developing generic scheduling frameworks that can balance computing efficiency and latency performance, such as Shenango[2] and SNAP[3].

All of this calls for intelligent solutions for the management of network functions in the context of radio access networks.

2. State of the art

2.1. Deep Reinforcement Learning-based radio schedulers

Deep Reinforcement Learning (DRL) is a popular technique for solving problems in dynamic scenarios as it doesn't require labeling new samples or training the model from scratch. In Comsa et al.[4], an actor-critic

[1] Xenofon Foukas and Bozidar Radunovic. 2021. Concordia: teaching the 5G vRAN to share compute. In Proceedings of the 2021 ACM SIGCOMM 2021 Conference (SIGCOMM '21). Association for Computing Machinery, New York, NY, USA, 580–596. https://doi.org/10.1145/3452296.3472894

[2] Ousterhout, Amy, Joshua Fried, Jonathan Behrens, Adam Belay, and Hari Balakrishnan. "Shenango: Achieving High CPU Efficiency for Latency-sensitive Datacenter Workloads." In NSDI, vol. 19, pp. 361-378. 2019.

[3] Michael Marty, Marc de Kruijf, Jacob Adriaens, Christopher Alfeld, Sean Bauer, Carlo Contavalli, Michael Dalton, Nandita Dukkipati, William C. Evans, Steve Gribble, Nicholas Kidd, Roman Kononov, Gautam Kumar, Carl Mauer, Emily Musick, Lena Olson, Erik Rubow, Michael Ryan, Kevin Springborn, Paul Turner, Valas Valancius, Xi Wang, and Amin Vahdat. 2019. Snap: a microkernel approach to host networking. In Proceedings of the 27th ACM Symposium on Operating Systems Principles (SOSP '19). Association for Computing Machinery, New York, NY, USA, 399–413. https://doi.org/10.1145/3341301.3359657

[4] I.-S. Comsa, A. De-Domenico, and D. Ktenas, “Qos-driven scheduling in 5g radio access networks-a reinforcement learning approach,” in GLOBECOM 2017-2017 IEEE Global Communications Conference, pp. 1–7, IEEE, 2017.

agent is proposed that selects the most appropriate scheduling algorithm to maximize Quality of Service (QoS) for users. The state includes the number of active users, arrival rate, Channel Quality Indicators (CQIs), and performance measure concerning QoS requirements of users. The reward function measures the actual impact of choosing a rule on meeting QoS objectives for users.

Naparstek et al. [5] proposed a lightweight multi-user DRL algorithm to address the problem of spectrum access. Users can choose only one channel to transmit per time slot, and an ACK signal is received if the transmission is successful. The state includes the history of transmissions on the selected channel and observations made by the user. The reward function measures the achievable throughput.

In Tan et al.[6], multiple access schemes are used to divide the time frame among Wi-Fi and LTE users. The authors proposed a DRL algorithm to find the proper splitting point according to feedback on the channel status for earlier time frames. The state includes idle slots, successful transmissions, and the reward function measures the transmission time given to LTE users without violating the throughput requirements of Wi-Fi users.

The authors of [7] focus specifically on the radio resource scheduling problem in the Medium Access Control (MAC) layer with the objective of jointly optimizing throughput and fairness. They proposed a DRL algorithm that works agnostically for various numerologies without retraining when the numerology changes. The objective is to select an active User Equipment (UE) from a candidate set and allocate available Resource Block Groups (RBG) to the chosen UE. The state includes UE eligibility, data rate, and fairness, and the reward function measures the throughput.

All of these works include channel conditions in their state space and consider UE throughput in their reward function.

2.2. CPU-aware radio schedulers

Dynamic variations in computational resource consumption by virtualized RAN can cause under-provisioning of resources, severely degrading network performance by, for example, causing radio frames not to be decoded on time. Some research has been done to determine the usage of computational resources by virtual RAN functions[8]. Bhaumik et al. [9]analyzed the computational resource usage of virtual RAN functions using a real experimental testbed but without considering the impact of Signal-to-Noise Ratio (SNR) on computational load. Rost et al. [10] proposed a model to determine the computational load by considering the

[5]O. Naparstek and K. Cohen, "Deep Multi-user Reinforcement Learning for Distributed Dynamic Spectrum Access," IEEE transactions on wireless communications, vol. 18, no. 1, pp. 310–323, 2018

[6] J. Tan, L. Zhang, Y.-C. Liang, and D. Niyato, "Deep reinforcement learning for the coexistence of laa-lte and wifi systems," in ICC 20192019 IEEE International Conference on Communications (ICC), pp. 1– 6, IEEE, 2019

[7] F. Al-Tam, N. Correia, and J. Rodriguez, "Learn to Schedule (LEASCH): A Deep Reinforcement Learning Approach for Radio Resource Scheduling in the 5G MAC layer," IEEE Access, vol. 8, pp. 108088–108101, 2020

[8] J. Bartelt, P. Rost, D. Wubben, J. Lessmann, B. Melis, and G. Fettweis, "Fronthaul and Backhaul Requirements of Flexibly Centralized Radio Access Networks," IEEE Wireless Communications, vol. 22, no. 5, pp. 105–111, 2015

[9] S. Bhaumik, S. P. Chandrabose, M. K. Jataprolu, G. Kumar, A. Muralidhar, P. Polakos, V. Srinivasan, and T. Woo, "CloudIQ: A Framework for Processing Base Stations in a Data Center," in Proceedings of the 18th annual international conference on Mobile computing and networking, pp. 125–136, 2012

[10] P. Rost, S. Talarico, and M. C. Valenti, "The Complexity–Rate Tradeoff of Centralized Radio Access Networks," IEEE Transactions on Wireless Communications, vol. 14, no. 11, pp. 6164–6176, 2015

impact of SNR and the applied Modulation and Coding Scheme (MCS) on computational resource consumption. They demonstrated that computational resource usage decreases when more robust MCS schemes are selected.

This idea is used by Bega et al. [11], where two main functions of the protocol stack are re-designed to avoid network performance degradation in virtualized RAN by avoiding computational outages and optimizing the MCS and resource block assignment on a per-frame basis. However, these latter works use models to provide their solutions, whereas a model-free solution based on Deep RL techniques is certainly better suited for this problem.

3. Use case – reliable RAN

Our focus in this work is on the Physical (PHY) and Media Access Control (MAC) layers of the 3GPP 5G-NR stack[12], which are responsible for a wide range of functionalities such as demodulation, decoding, and radio resource scheduling for UEs. Fig. 1 illustrates our model, and we provide the notation used in this work in Table 1.

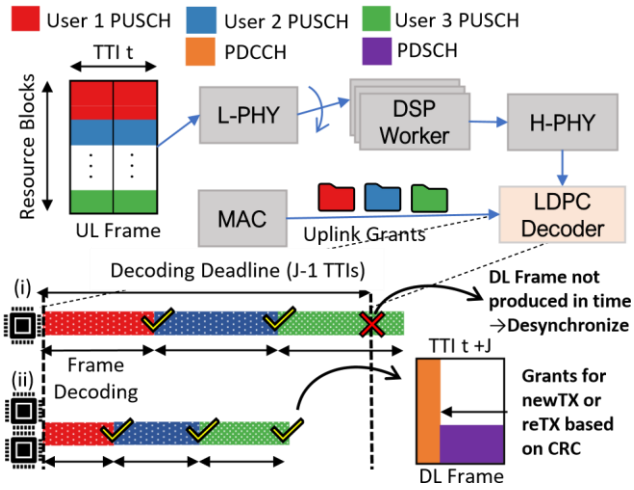

Figure 1. Decoding in 5G-NR. i) On congested CPUs: A deadline violation occurred and the DL frame was not generated in time. ii) On not congested CPUs: The grant carrying information in new transmission (newTX) or retransmission (reTX) is packed in the PDCCH of the DL Frame.

Table 1 Notation Table

Short	Expansion	Variable	Description	Set
UE	User Equipment	*u*	User	U
MCS	Modulation Coding Scheme	*mu*	User's *u* MCS	M
PRB	Physical Resource Blocks	*nu*	User's *u* PRBs	N
TBS	Transport Block Size	*bu*	User's *u* TBS	T

[11] D. Bega, A. Banchs, M. Gramaglia, X. Costa-Perez, and P. Rost,´ "CARES: Computation-Aware Scheduling in Virtualized Radio Access Networks," IEEE Transactions on Wireless Communications, vol. 17, no. 12, pp. 7993–8006, 2018

[12] 3GPP, "3rd Generation Partnership Project; 5G; NR; Overall description; Stage-2 ." 3GPP TS 38.300 Release 18, May 2022

TTI	Transmission Time Interval	*t*		

3.1. System Model

Our system comprises of a base station, or gNB, and a set Uc of user equipments (UEs) u that utilize the communication services provided by the gNB. The UEs transmit their data to the gNB, which allocates radio resources to them based on predefined scheduling policies. These policies specify the number of Physical Resource Blocks (PRBs) and Modulation and Coding Schemes (MCS) to be used for transmission, details of which we present later.

During each Transmission Time Interval (TTI), the gNB assigns a specific number of PRBs and an MCS to each UE for both Uplink (UL) and Downlink (DL) transmissions. We primarily focus on the UL pipeline, as it imposes a significant computing burden on the gNB [13]. Moreover, it is a critical factor to consider for vRAN systems..

3.2. Physical Resource Blocks

The base station, or gNB, manages the total available bandwidth which is partitioned into N blocks, known as Physical Resource Blocks (PRBs). In both 5G-NR and LTE systems, an Orthogonal Frequency-Division Multiple Access (OFDMA) scheme is used for data transmission, where data is encoded into a portion of the N available blocks allocated to a UE.

In 5G-NR systems, the number of available PRBs, N, is variable and depends on the total available bandwidth and the width of each PRB. For instance, in a scenario with 10 MHz bandwidth and 180 kHz-wide PRBs with a 20 kHz guard band, there are N=50 available PRBs during each Transmission Time Interval (TTI).

At each TTI, the gNB decides on how to assign resources to UEs. Specifically, let nu be the number of PRBs assigned to UE u for their Uplink (UL) communication with the gNB. The total allocated PRBs to all users must not exceed the total available PRBs in the system, that is, the sum of nu for all u in the set of UEs must be less than or equal to N. We focus on the UL pipeline since it creates a significant computing load for the gNB and is crucial for vRAN systems.

3.3. Modulation and Coding Schemes

Each user's Modulation Coding Scheme (MCS) mu belongs to a set of M elements, where each element captures a number of bits that will be encoded in each symbol transmitted by user u. For 5G-NR, 3GPP defines 29 MCS indexes, ranging from the lowest BPSK capable of carrying 1 bit/symbol, up to 256-QAM that can carry 8 bits/symbol.

Combined with MIMO techniques, these MCSs provide the Gbps data rates obtained by 5G. The MCS that will be used by the UEs depends on the estimated channel conditions, which are upper bounded by Shannon's law. Thus, in practice, not all MCS indices are always available.

[13] S. Bhaumik, S. P. Chandrabose, M. K. Jataprolu, G. Kumar, A. Muralidhar, P. Polakos, V. Srinivasan, and T. Woo, “CloudIQ: A Framework for Processing Base Stations in a Data Center,” in Proceedings of the 18th annual international conference on Mobile computing and networking, pp. 125–136, 2012

3.4. Radio Scheduling Procedure

The Radio Resource Scheduling Procedure in 5G-NR UL involves a centralized agent, such as the network operator in the gNB, making decisions regarding the user u granted data for the UL channel at each Transmission Time Interval (TTI). The procedure involves periodic transmission of a Buffer State Report (BSR) by each user u to the gNB, which contains the user's UL demand, i.e., the size (in bytes) of the total data pending in the user equipment (UE) u stack. Based on this report, the gNB schedules the user u transmission and responds with a transmission grant that includes two decisions: (i) the number nu of Physical Resource Blocks (PRBs) that are assigned to user u, and (ii) the Modulation and Coding Scheme (MCS) mu that u must use. The choice of the granted user $u \in U$ and the selection of nu and mu depend on the specific scheduling algorithm implemented by the gNB, which uses information such as the user's data size or their channel conditions, such as their Signal-to-Noise Ratio (SNR), which can limit the communication.

The selection of MCS and PRBs eventually leads to a specific amount of data that can be transmitted by a UE, called Transport Block (TB), with a specific Size (TBS) bu that accounts for the user data, the size of the extra information such as Cyclic Redundancy Checks (CRCs), and filler bits. Given the discrete number of PRBs and MCS, the possible values of TBS can be found using a simple deterministic procedure.

3.5. Radio Scheduling Procedure in VRAN systems

The problem of Resource Scheduling in vRAN involves maximizing the system throughput in the communication between end-users and the gNB, considering possible impairments due to the cloud infrastructure. A deep learning approach is considered to tackle the problem. According to the 3GPP 5G-NR specification, all the tasks related to uplink (UL) frame decoding at the Physical Layer must be completed before the next downlink (DL) frame is sent. There is a hard deadline for the completion of the decoding tasks, which are by far the most computationally expensive of the full protocol stack. The deadline is configurable in 5G-NR and is set to J-1 Transmission Time Intervals (TTIs), meaning that the frame decoding result has to be ready before the ACK/NACK transmission that will take place in the J-th TTI after the reception of the original frame. The default values for the TTI length and J impose a deadline of 3ms to complete demodulation.

4. Conclusion

In this document, we have presented a comprehensive analysis of the radio resource scheduling procedure in virtualized Radio Access Networks (vRANs), with a focus on optimizing the achieved throughput by taking into account the status of the underlying cloud infrastructure. We have provided a detailed description of the 5G-NR UL radio resource scheduling procedure and how it enables the centralized agent, such as the network operator, to make informed decisions on granting data transmission to the user equipment (UE). Furthermore, we have identified the critical challenges in vRANs, including the integration of hardware accelerators and the optimal interplay between the cloud infrastructure and the vRAN software. To address these challenges, we proposed a deep learning approach that can effectively maximize the system throughput while considering possible impairments due to the cloud infrastructure. Our study highlights the potential benefits of integrating machine learning techniques in the design of vRAN systems, which can lead to significant improvements in the overall network performance and user experience.