Deep Reinforcement Learning for Orchestrating Cost-Aware Reconfigurations of vRANs

Murti, Fahri Wisnu; Ali, Samad; Iosifidis, George; Latva-aho, Matti

DOI
10.1109/TNSM.2023.3292713

Publication date
2024

Document Version
Final published version

Published in
IEEE Transactions on Network and Service Management

Deep Reinforcement Learning for Orchestrating Cost-Aware Reconfigurations of vRANs

Fahri Wisnu Murti, Samad Ali, George Iosifidis, and Matti Latva-aho, *Senior Member, IEEE*

***Abstract*—Virtualized Radio Access Networks (vRANs) are fully configurable and can be implemented at a low cost over commodity platforms to enable network management flexibility. In this paper, a novel vRAN reconfiguration problem is formulated to jointly reconfigure the functional splits of the base stations (BSs), locations of the virtualized central units (vCUs) and distributed units (vDUs), their resources, and the routing for each BS data flow. The objective is to minimize the long-term total network operation cost while adapting to the varying traffic demands and resource availability. In the first step, testbed measurements are performed to study the relationship between the traffic demands and computing resources, which reveals high variance and depends on the platform and its load. Consequently, finding the perfect model of the underlying system is non-trivial. Therefore, to solve the proposed problem, a deep reinforcement learning (RL)-based framework is proposed and developed using model-free RL approaches. Moreover, the problem consists of multiple BSs sharing the same resources, which results in a multi-dimensional discrete action space and leads to a combinatorial number of possible actions. To overcome this curse of dimensionality, action branching architecture, which is an action decomposition method with a shared decision module followed by neural network is combined with Dueling Double Deep Q-network (D3QN) algorithm. Simulations are carried out using an O-RAN compliant model and real traces of the testbed. Our numerical results show that the proposed framework successfully learns the optimal policy that adaptively selects the vRAN configurations, where its learning convergence can be further expedited through transfer learning even in different vRAN systems. It also offers significant cost savings by up to 59% of a static benchmark, 35% of Deep Deterministic Policy Gradient with discretization, and 76% of non-branching D3QN.**

***Index Terms*—Radio access networks (RANs), network virtualization, O-RAN, orchestration, deep reinforcement learning, D3QN, action branching.**

I. Introduction

A. Motivation

VIRTUALIZATION has become one of the most promising technologies for accommodating the increased service demands with diverse requirements at a reasonable cost in cellular networks [1]. The latest effort of this idea is virtualizing the radio access networks (vRANs) by replacing the hardware-based legacy RANs with softwarized RANs [2], [3], [4]. Incorporated with Open RAN, vRANs can be fully configurable and deployed across heterogeneous platforms such as commodity servers and small embedded devices. Another exciting feature of vRANs is that it enables the baseband functions (BBU) of each base station (BS) to be disaggregated, hosted at the virtualized distributed units (vDUs) and central units (vCUs), and executed as virtual machine (VM) instances or light-weight containers over geo-distributed locations. This paradigm shift brings unprecedented flexibility to RAN operations, mitigates vendor lock-in, offers fast deployment and potentially reduces operational expenses [4]. Therefore, it is not surprising that many standardization bodies envision the virtualization for their future RANs, such as O-RAN [5] and 5G+ RAN [6].

Nevertheless, the expansive deployment of vRANs is still hindered by complex configuration options, which introduce new network management challenges in deploying cost-efficient vRAN configurations while serving the traffic demands. In particular, the operators need to decide the *functional splits* of the BSs to determine which BS functions are deployed at the vDUs and which are at the vCUs. Each choice of these splits has a different delay requirement, consumes different computing resources for the vDUs/vCUs, and generates a different data load over the xHaul links.[1] Moreover, the vDUs/vCUs are executed on top of commodity platforms as VM instances or containers; hence, the operators need to allocate the *virtualized resources* (e.g., CPUs, memory) for them. There are also several candidate deployment locations for each vDU/vCU, possibly with different hosting machines, and this creates the *placement problem* in determining their optimal locations and platforms. At the same time, each placement location is associated with different eligible *routing paths* to transfer the data flow of the BSs, which incur particular delays and costs. Consequently, these issues create a challenging coupling among the BS splits, placement and allocated resources for the vDUs/vCUs, and routing for each BS data flow.

Meanwhile, the suitability of the vRAN configurations is highly affected by the network properties such as traffic demands and resource availability (e.g., computing and xHaul link capacity) [7], which might change over time, often in

Manuscript received 1 December 2022; revised 1 May 2023; accepted 29 June 2023. Date of publication 5 July 2023; date of current version 7 February 2024. This research has been supported by the Academy of Finland, 6G Flagship program under Grant 346208. Fahri Wisnu Murti also would like to acknowledge the support of Nokia Foundation. The associate editor coordinating the review of this article and approving it for publication was R. Riggio. *(Corresponding author: Fahri Wisnu Murti.)*

Fahri Wisnu Murti, Samad Ali, and Matti Latva-aho are with the Centre for Wireless Communications, University of Oulu, 90570 Oulu, Finland (e-mail: fahri.murti@oulu.fi).

George Iosifidis is with the Software Technology, Delft University of Technology, 2600 AA Delft, The Netherlands.

Digital Object Identifier 10.1109/TNSM.2023.3292713

[1]The paths connecting a core network (EPC) to vCUs, vCUs to vDUs, and vDUs to radio units (RUs) are defined as backhaul (BH), midhaul (MH), fronthaul (FH), respectively, and the integration of these elements is called Crosshaul/xHaul transport network.

an unpredictable fashion.[2] Thus, deploying static configurations for a long time might result in resource overprovisioning or even declined traffic demands. Resource overprovisioning occurs when the allocated resources are higher than the actual resource utilization. The declined demands can be triggered by insufficient allocated resources (underprovisioning) and constraint violation. And these can render substantial performance degradation and high operating expenditures. Therefore, it is essential to dynamically select the vRAN configurations to adapt to varying traffic demands and resource availability.

On the other hand, orchestrating the dynamic configurations of vRANs is a non-trivial endeavor. The reconfiguration decisions must be enforced before the actual traffic demands of the BSs are observed. Albeit reconfiguring the vRAN system at runtime is practically possible [9], it might also induce additional costs and disrupt network operations during the live migration of the VM instances. Consequently, any reconfiguration activity needs to be performed prudently to ensure that it is beneficial both in terms of cost and performance. However, designing such an intelligent approach also has technical issues since the software-based vRAN system substantially differs from hardware-based legacy RANs. The takeaway from our testbed measurements (details in Section V) and prior experimental studies (cf., [10], [11]) is that, unlike legacy RANs, the underlying system of vRANs is complex, poly-parametric and has platform-dependent performance. Hence, adopting traditional control policies, which needs perfect knowledge of the underlying system to model and solve the problems, is unrealistic in practice.

Motivated by the challenges above and our measurement insights (details in Section V), we propose and study *a fresh vRAN reconfiguration problem*, where it jointly reconfigures the splits of the BSs, resources and locations of the vDUs and vCUs, and the routing of each BS data flow to minimize the long-term total network operation cost. The key idea is to model this problem as reinforcement learning (RL) and develop *a learning-based framework*, namely **L**earning-based **A**utomated **R**econfiguration for **v**RANs (LARV), to solve the problem with minimal assumptions about the system.

B. Contributions and Methodology

We firstly build a prototype implementing the centralized-RAN (C-RAN) system using software-based srsRAN [12] in two different platforms to collect measurements regarding the relations between traffic demands and resource utilization. The findings reveal that the relations vary with the demands and, importantly, have high variance and dependence on the platform, platform load,[3] and many latent factors. These inhibit adopting general assumptions of the underlying system (e.g., linear) and traditional mathematical tool-based policies. Then, we propose a new cost model accounting for resource overprovisioning, instantiation and reconfiguration, and the declined traffic demands, representing the virtualized resource management in vRANs. This model also considers different computing and routing costs for each split and platform location. Further, we model our vRAN system following the latest proposal of O-RAN architecture [5]. We consider a vRAN system with multiple BSs and define its operation as a time-slotted system, where each slot has arbitrary incoming traffic demands and resource availability. At each time slot, LARV takes an action that selects the vRAN configurations, then reconfigures the system when the selected are different from the last configurations or preserves them if the selected configurations are the same. LARV expects to receive a reward signal from the system that assesses the quality of each selected action. This sequential decision-making is formulated as Markov decision process (MDP), which is also an RL problem.

In our solution, LARV is developed using model-free RL with deep neural network architecture. LARV considers the vRAN system as a black-box environment and does not make any particular assumptions about the underlying system state and state transition probability distribution. Since the formulated RL problem has a semi-continuous state space and discrete action space, we propose a Dueling Double Deep Q-network (D3QN)-based approach [15], in which the learning step is based on Double Q-learning [16]. However, the system has multiple BSs that share the same resources with highly coupled configuration decisions. As a result, the RL formulation renders a multi-dimensional action space, which exhibits combinatorial growth of the number of possible actions. In order to overcome the curse dimensionality, the proposed D3QN is incorporated with action branching [17], an action decomposition method that decomposes the multi-dimension action into sub-actions and utilizes shared decision module followed by neural network branches. However, the initial action branching proposed in [17] focused on sub-actions with the same dimensional size, which can not be directly applied to our problem. Here, we adapt it; hence each sub-action dimension can vary but still exhibits a linear growth of the total neural network outputs (estimated actions) with the increase of action dimensionality while maintaining the shared decision.

We conduct a battery of tests using an O-RAN compliant model and real traces collected from the testbed. We evaluate the training behavior and long-term total network cost during online operation under various scenarios. Our numerical results reveal that LARV successfully learns the optimal policy to select an action that controls the vRAN configurations, where its learning convergence can be accelerated via transfer learning even in different vRAN systems. Moreover, LARV offers considerable cost savings by up to 59% of a static benchmark, 35% of Deep Deterministic Policy Gradient (DDPG) with discretization, and 76% of distributed non-branching D3QN. Our contributions can be summarized:

- We propose and study a new vRAN reconfiguration problem, where it jointly reconfigures splits of the BSs, resources and locations of the vDUs/vCUs, routing for each BS flow.
- We carefully model our vRAN system based on the latest proposals of O-RAN architecture and propose a

[2]This is particularly common for resource availability/costs in shared infrastructures or traffic and channel conditions in small cell networks [8].

[3]The relations heavily rely on the types of platforms that host the BBU. It also depends on platform load (e.g., when vRAN workload shares the same platform with other applications or workloads such as edge computing, data analytics, etc.); see Section V and [10], [13], [14] for details.

comprehensive cost model. The model takes resource overprovisioning, instantiation and reconfiguration and the declined demands costs into account. It also captures platform/split-dependent computing and routing costs.
- We develop a learning-based framework to solve the proposed vRAN reconfiguration problem. It is tailored from D3QN and an action branching architecture to tackle the multi-dimensional and large action space inherited from our RL problem with linear growth of the neural network outputs.
- We conduct extensive trace-driven simulations and analyze the performance of LARV under various scenarios during the training process and online operation.

The rest of this paper is organized as follows. Section II discusses our contributions with respect to prior works. In Section III, the architecture background and model used for our vRAN system are presented. The reconfiguration problem is also formulated in this section, including the raised trade-offs. In Section IV, we discuss how to design the proposed learning algorithm. The detailed experiment setups, testbed measurement insights, and simulation results are presented in Section V. Finally, our paper is concluded in Section VI.

II. Related Work

Recent works have studied various vRAN orchestration problems, and we can classify them into *i)* those that rely on models to optimize the configurations and *ii)* model-free approaches that utilize offline training data and *iii)* RL methods. The examples of the first point include [18], [19] that optimize the vRAN functional splits with multi-access edge computing (MEC) services, [7] that considers the functional split problem with multiple servers for hosting the vCUs, and [20] that further expands it to several candidate servers to place the vCUs/vDUs. Albeit they have optimized various configurations in vRANs, they aimed for offline network designs, and the implication of varying conditions from traffic demands and resource availability is still not examined. The studies of model-based approaches that consider varying conditions include altering the functional splits at runtime to maximize the users' throughput [21] and revenue [22] and to minimize inter-cell interference and FH utilization [23]. Another example in [24] aimed to control radio/computing scheduling to maximize the served traffic subject to a BS computing capacity. However, they [21], [22], [23], [24] still did not study where to place and how much the allocated resources are for the vDUs/vCUs, although these configurations play crucial role in a vRAN system. Moreover, such model-based approaches can be impractical as they heavily rely on fine-tuning models for specific scenarios and underlying system assumptions. And a vRAN system is network and platform-dependent, where the models can be unknown in practice.

On the other hand, model-free approaches employing machine learning (ML) have been increasingly popular in tackling complex problems in mobile networks. Particularly, approaches that employ function approximation of performance metrics, e.g., via neural networks, can offer satisfactory performance amidst many unknown system parameters [25]. For instance, the authors in [26] have developed a deep supervised learning framework for allocating radio resources and functional split for each user. Such supervised learning can deliver well-achieved performance as long as there are high-quality labeled datasets, e.g., optimal labels. However, the optimal labels are often not be available in vRAN problems. Hence, those that do not require labeled datasets, such as contextual bandit and full RL formulations, can be leveraged. The authors in [10] have tailored a deep learning-based framework to solve the contextual bandit problem of managing the interplay between computing and radio resources. The other contextual bandits in [11] and [27] utilize a data-efficient algorithm, Bayesian online learning for an energy-aware BS in a vRAN system. These approaches offer remarkable performance with the condition that the current context observation must not be affected by the previous actions, i.e., it only includes exogenous parameters.

Otherwise, a full RL formulation is required when the current observation, e.g., state, is influenced by the previous actions. Recent work in [28] has brought the importance of a model-free RL formulation by utilizing Q-learning and SARSA algorithms to optimize the functional split selections for an energy-efficient O-RAN. However, when the state-action space of the RL problem is large, such approaches become inefficient. Therefore, a deep RL paradigm can be utilized to tackle such issue by using neural network architecture to approximate the state-action function. Some interesting examples are [29] and [30] that have developed xApps for controlling RAN slicing, scheduling and online model training using the Proximal Policy Optimization algorithm. In [31], the authors also have solved the functional split problem by proposing a chain rule-based stochastic policy and approximate it with sequence-to-sequence model. Our recent work in [32] has proposed an RL-based framework using a combination of Deep Q-Network (DQN) and a regressor to dynamically reconfigure the functional split and its required computing resources. However, it was still limited to a single BS and did not include computing and link resource sharing among the BSs.

Although the mentioned works have solved various adaptive vRAN orchestration problems, they mainly focused on controlling functional splits (e.g., [21], [22], [23], [24], [28]), RAN slicing (e.g., [29], [30]) and radio/computing scheduling (e.g., [10], [11], [26], [27], [28], [29], [30]). On the other hand, the joint reconfiguration between functional splits of the BSs, the virtualized resource allocation and placement for the vCUs/vDUs over geo-distributed cloud platforms, and the routing, along with the impacts of altering such configurations at runtime, are *hitherto unexplored*. Here, we aim to fill a gap by tackling this reconfiguration problem using model-free RL that makes minimal assumptions about the system. Since the problem also consists of multiple BSs with highly coupled configurations, the RL formulation renders a dimensional explosion in the state space and action space, making the available vRAN orchestration frameworks unsuitable. To solve this challenging dimensionality issue, we develop LARV, a novel

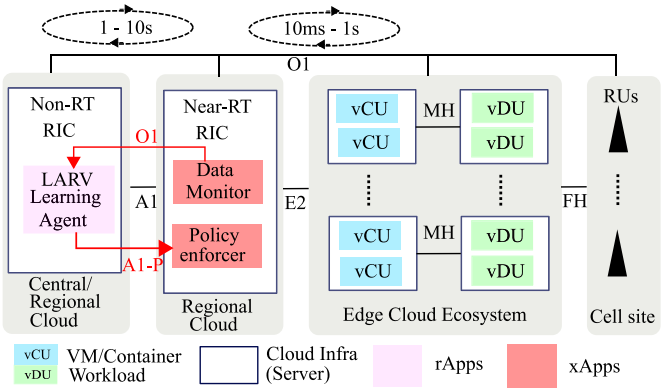


Fig. 1. O-RAN compliant system architecture adopted in our model.

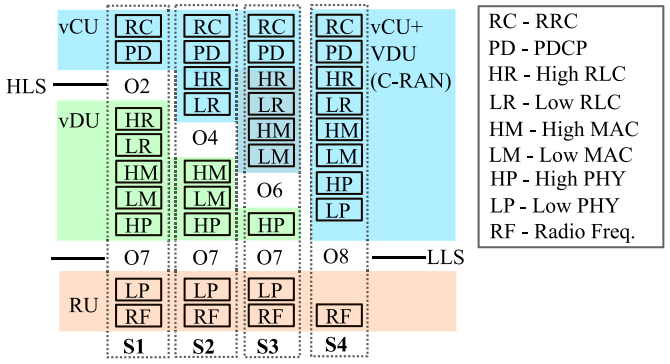


Fig. 2. The functional splits applied in our vRAN model. S1, S2 and S3 are envisioned in O-RAN architecture proposals, while S4 are legacy C-RAN.

vRAN orchestration framework based on deep RL, from the incorporation of action branching with D3QN.

III. System Model and Problem Formulation

A. Background and Model

We model our vRAN system following the latest proposals of O-RAN architecture [5], where the high-level architecture is illustrated in Fig. 1. The model adopts O-RAN key principles that include disaggregation, virtualization, open interfaces, and intelligent control [33]. The protocol stacks (or functions) of each BS can be disaggregated through the functional split and, further, virtualized as the vCU and vDU (connected to an RU). Hence, a BS corresponds to 4G eNodeB or 5G gNodeB comprising a vCU, vDU, and RU. The vCU and vDU can be executed as VM instances or containers across geo-distributed edge cloud infrastructures, which may share with other workloads. Then, the intelligent control is realized through RAN Intelligent Controllers (RICs), which can run routine optimization and orchestration through closed-loop control. O-RAN has specified two RICs: i) Non-Real-Time (Non-RT) RIC and ii) Near-Real-Time (Near-RT) RIC. The Non-RT RIC, which integrates with the network orchestrator, operates on a time scale longer than 1 s, while the Near-RT RIC operates with a time scale between 10 ms and 1s. The Non-RT RIC supports applications, called rApps, that support RAN optimization and operations such as policy guidance, configuration management, etc. While the Near-RT RIC includes applications called xApps that can be used to perform radio resource management. Then, LARV is to be implemented in the learning agent as an rApp in the Non-RT RIC in the system orchestrator of O-RAN and enforces a policy at every period of $n = 1, \ldots, N$ to control the reconfigurations of BSs. The optimal policy at every time n depends on the input observation (state), which is provided at the beginning of each period by the BSs via the O1 interface.

Next, we illustrate the functional split options used in our model in Fig. 2 and present their requirements in Table I. As suggested by O-RAN [5], we consider Option 7.x (O7) and Option 8 (O8) for the Low Layer Split (LLS) between the vDU and RU. The High Layer Split (HLS) between the vCU and vDU can use Option 2 (O2), which is currently the most feasible split to be implemented. We also consider Option 4 (O4) and Option 6 (O6), which have been well standardized [5], [6]

TABLE I
The Functional Split Options and Their Requirements Based on 3GPP Nomenclature When the Traffic Demand Is λ Gbps. The Requirements Are Tailored by Following Settings: 100 MHz Bandwidth, 256 QAM, 32 Antenna Ports and 8 MIMO Layers. The Achievable Data Rate Is Up to 4 Gbps

	Split Point	Load	Max	Delay Req.
O1	RRC - PDCP	λ	4	10 ms
O2*	PDCP - High RLC	λ	4	10 ms
O3	High RLC - Low RLC	λ	4	10 ms
O4*	Low RLC - High MAC	λ	4	1 ms
O5	High MAC - Low MAC	λ	4	1 ms
O6*	Low MAC - High PHY	1.02λ+0.5	4.13	0.25 ms
O7†	High PHY - Low PHY	10.1	10.1	0.25 ms
O8†	Low PHY - RF	157.3	157.3	0.25 ms

Note: * is applied options for HLS and † is applied options for LLS. The data load is in Gbps.

and experimentally validated [21], to encourage further RAN flexibility. Therefore, following HLS and LLS, we denote four choices of functional splits: *Split 1 (S1)* implements O2 for the HLS and O7 for the LLS; *Split 2 (S2)* uses O4 for the HLS and O7 for the LLS; *Split 3 (S3)* adopts O6 for the HLS and O7 for the LLS; and *Split 4 (S4)* is the legacy C-RAN system, which implements Option 8 (O8), i.e., all the BS functions are executed as an integrated vDU/vCU except RF functions (at the RU). We define a set of these four possible splits as $\mathcal{I} = \{\text{S1}, \text{S2}, \text{S3}, \text{S4}\}$.

We consider a vRAN system with K BSs, where the functions of each BS-k can be disaggregated and hosted at vCU-k, vDU-k and RU-k. The vDUs are executed at far-edge cloud servers (FSs) while the vCUs are at edge cloud servers (ESs).[4] We model a packet-based vRAN as a graph of $G = (\mathcal{V}, \mathcal{E})$, where the set of physical nodes $\mathcal{V}$ includes the subsets: $\mathcal{K} = \{1, \ldots, K\}$ of RUs, $\mathcal{L} = \{1, \ldots, L\}$ of FSs, $\mathcal{M} = \{1, \ldots, M\}$ of ESs, EPC (index 0), and routers. These nodes are connected through a set of links $\mathcal{E}$, where each link $(i, j) \in \mathcal{E}$ has a data transfer capacity c_{ij} (Gbps). We denote $\mathcal{P}_k$ as a set of paths connecting EPC to RU-k and consider the data flow for each BS is unsplittable. We focus on the downlink, but it is not

[4]FSs are the candidate platforms and locations to execute VM instances of the vDUs. Similarly, ESs are the candidate platforms and locations for the vCUs. We also consider ESs for the candidate platforms to host an integrated vDU/vCU in C-RAN. ESs are typically located at more centralized locations, while FSs are co-located or near the RUs.

TABLE II
KEY VARIABLES AND PARAMETERS USED IN OUR MODEL

Descriptions	Notations
The traffic demand (split) of BS-k	λ_k^n (i_k^n)
Allocated flavors (actual resource utilization) for vDU-k/vCU-k	x_k^n/y_k^n $(\hat{x}_k^n/\hat{y}_k^n)$
Locations of vDU-k and vCU-k	z_k^n, ζ_k^n
Maximum computing capacity of FS-l and ES-m	H_l, $\hat{H}_m$
A connecting path of EPC→ES-m, ES-m→FS-l, ES-m→RU-k, FS-l→RU-k	p_{0m}, p_{ml}, p_{mk}, p_{lk}
Incurred delay of path p_{0m}, p_{ml}, p_{mk}, p_{lk}	$d_{p_{0m}}, d_{p_{ml}}, d_{p_{mk}}, d_{p_{lk}}$
HLS and LLS delay requirement for split i	d_i^H, d_i^L
The routing for BS-k	$p \in \mathcal{P}_k$
Data flow with split i via routing $p \in \mathcal{P}_k$ (FH, MH , BH)	$(r_{p,i}^{\text{FH},n}, r_{p,i}^{\text{MH},n}, r_{p,i}^{\text{BH},n})$

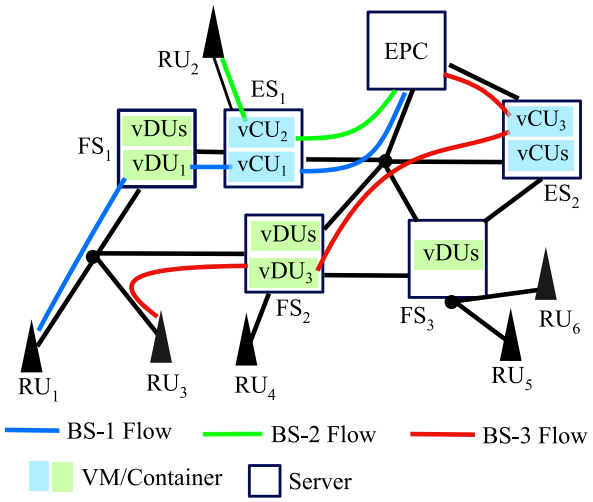


Fig. 3. The functions of each BS can be split between the vCU, vDU and RU. For BS-1, vDU-1 and vCU-1 are executed at FS-1 and ES-1, respectively. However, BS-2 implements C-RAN (S4); hence, the integrated vDU/vCU are executed only at ES-1 (e.g., links to and instances at FSs are not activated). Then, vDU-3 and vCU-3 of BS-3 are hosted at FS-2 and ES-2, respectively.

limited and can easily be extended for uplink. The data flow for each BS will be transferred from EPC to RU-k through a path $p := \{(0, i_1), (i_1, i_2), \ldots, (i_k, k) : (i, j) \in \mathcal{E}\} \in \mathcal{P}_k$. Since this path might pass through FSs and ESs before reaching each RU, let us denote p_{0m}, p_{ml}, p_{mk}, and p_{lk} as a path connecting EPC→ES-m, ES-m→FS-l, ES-m→RU-k, and FS-l→RU-k, respectively. Based on the selected split, the data flow of each BS-k passes through $p := p_{0m} \cup p_{ml} \cup p_{lk} \in \mathcal{P}_k$ (EPC → ES-m → FS-l → RU-k) if activating S1, S2 and S3. Otherwise (e.g., S4/C-RAN), the flow passes through $p := p_{0m} \cup p_{mk} \in \mathcal{P}_k$ (EPC → ES-m → RU-k without using FSs). Each path has a total delay defined as $d_p, d_{p_{0m}}, d_{p_{ml}}, d_{p_{mk}}$ and $d_{p_{lk}}$; and they must respect the delay requirements of the split as described in Table I. We compute each p_{0m}, p_{ml}, p_{mk} and p_{lk} with the shortest path method. Fig. 3 shows an example of our model.

We use the term *flavor*[5] to define the available choices for allocating the virtualized computing resources. Let us introduce $\mathcal{X}$ as a set of available flavors for the vDUs and vCUs. Then, we select a flavor $x_k \in \mathcal{X}$ and $y_k \in \mathcal{X}$ that determine the reserved resources for each vDU-k (in FSs) and vCU-k (in ESs). Each FS-l has physical computing capacity H_l, respectively $\hat{H}_m$ for ES-m, which bound the aggregate allocated resources (accordingly, the flavors that can be selected) of the vDUs and vCUs for each location. The key notations used in our model are summarized in Table II.

[5]This term is carried out from OpenStack (https://www.openstack.org/) to reserve the amount of virtual CPU, memory, and storage capacity for a VM instance. This term is typically used to calculate the billing units to charge the amount of monetary cost. Similar terms are also used in other cloud services such as AWS and Azure. Here, we focus on the CPU resources as they are the most affected performance by the traffic demands.

B. Problem Formulation

We model the vRAN operation as a time-slotted system. Given an incoming sequence of possibly-different traffic demands and resource availability, we aim to design a policy (strategy) of an agent that controls the vRAN configurations at each time slot, which includes the splits of the BSs, flavors and locations of vDUs and vCUs, and the routing for each BS data flow, to minimize the long-term total network operation cost. This sequential decision problem is formulated as MDP, specified by a tuple $\{\mathcal{S}, \mathcal{A}, P, r\}$. At every time slot n, the agent observes a state from the state space $s^n \in \mathcal{S}$, then takes an action that selects the vRAN configurations from the action space $a^n \in \mathcal{A}$. Following each enforced action, the agent expects to receive a reward signal $r(s^n, a^n)$ as feedback from the environment (vRAN system). Since the state may not be stationary, we define $P(s^{n+1}|s^n, a^n)$ as the state transition probability that maps a state-action pair at time step n into the distribution of next states. And we take no assumption about it. The formulated problem is also naturally an RL problem, and we describe it as follow.

1) Action: We introduce $i^n := \{i_k^n \in \mathcal{I} : k \in \mathcal{K}\}$ as control variables to select the functional splits that decide which functions of the BSs to be placed at the vDUs and vCUs. The selection of the flavors that allocates the resources for the vDUs and vCUs is determined using control variables $x^n := \{x_k^n \in \mathcal{X} : k \in \mathcal{K}\}$ and $y^n := \{y_k^n \in \mathcal{X} : k \in \mathcal{K}\}$, respectively. We can determine the locations of vDUs over FSs and vCUs over ESs by $z^n := \{z_k^n \in \mathcal{L} : k \in \mathcal{K}\}$ and $\zeta^n := \{\zeta_k^n \in \mathcal{M} : k \in \mathcal{K}\}$. The routing paths to transferred the data flow of each BS is selected through variables $p^n := \{p^n \in \mathcal{P}_k : k \in \mathcal{K}\}$. Since routing variable $p^n \in \mathcal{P}_k$ depends on the placement of the vDU and vCU, we can determine $p := \{p_{0m} \cup p_{ml} \cup p_{mk} \cup p_{lk}\} \in \mathcal{P}_k$ directly from i_k^n, z_k^n and ζ_k^n. For instance, if BS-5 with $i_5^n :=$ S1 decides $z_5^n := 1$ and $\zeta_5^n := 2$, then the selected path becomes

$p \coloneqq \{p_{0,2} \cup p_{2,1} \cup \cup p_{1,5}\} \in \mathcal{P}_5$ with the transferred data flow EPC→ES-2→FS-1→RU-5. Therefore, we can treat $p^n \in \mathcal{P}_k$ as part of the environment. Then, we formalize the action at time slot *n* as:

$$a^n = \{i^n, x^n, y^n, z^n, \zeta^n\} \in \mathcal{A}, \ \mathcal{A} \coloneqq \left\{\mathcal{I} \times \mathcal{X}^2 \times \mathcal{L} \times \mathcal{M}\right\}^{|\mathcal{K}|}, \quad (1)$$

where this action is taken from the action space $\mathcal{A}$ of a finite set that includes all possible pairs of the reconfiguration control decisions from all the BSs.

2) State: The state observation at each time slot *n* of the RL problem consists of *(i)* The incoming traffic demands of the BSs $\lambda^n \coloneqq \{\lambda_k^n \in \mathbb{R}_+ : k \in \mathcal{K}\}$ (Gbps); *(ii)* the previous deployed splits $i^{n-1} \coloneqq \{i_k^{n-1} \in \mathcal{I} : k \in \mathcal{K}\}$; *(iii)* the previous allocated resources (flavors) for the vDUs $x^{n-1} \coloneqq \{x_k^{n-1} \in \mathcal{X} : k \in \mathcal{K}\}$ and *(iv)* vCUs $y^{n-1} \coloneqq \{y_k^{n-1} \in \mathcal{X} : k \in \mathcal{K}\}$; and *(v)* the previous deployed locations of each vDU-*k* over FS $z^{n-1} \coloneqq \{z_k^{n-1} \in \mathcal{L} : k \in \mathcal{K}\}$ and *(v)* each vCU-*k* over ES $\zeta^{n-1} \coloneqq \{\zeta_k^{n-1} \in \mathcal{M} : k \in \mathcal{K}\}$. It provides *time dynamic* of our variable interests: *(i)* the demand that needs to be served by each BS; *(ii)* the current active splits of the BSs; *(iii)* the availability of resources for each vDU and *(iv)* vCU; and *(v)* the availability to execute each vDU at FS and *(vi)* each vCU at ES. Then, the state observation at time slot *n* can be denoted:

$$s^n \coloneqq \left\{\lambda^n, i^{n-1}, x^{n-1}, y^{n-1}, z^{n-1}, \zeta^{n-1}\right\} \in \mathcal{S}, \quad \mathcal{S} \coloneqq \left\{\mathbb{R} \times \mathcal{I} \times \mathcal{X}^2 \times \mathcal{L} \times \mathcal{M}\right\}^{|\mathcal{K}|}. \quad (2)$$

The state space $\mathcal{S}$ is semi-continuous because it contains continuous parameters $\lambda_k^n \in \mathbb{R}_+, \forall k \in \mathcal{K}$ from the traffic demands. It is exogenous parameter, i.e., it is not affected by the action, but it provides contextual information about the users' needs. The other points are discrete parameters and provide the network state information, which are highly affected by the deployed configurations from the last action. This state information is provided as input to the learning agent through the O1 interface. The state can be extended to other relevant key performance measurements; however, the state space of the RL problem also expands.

3) Reward & Policy: Our reward function is calculated from the incurred total network operating cost. The source of monetary costs comes from the computing cost to execute the BS functions, the virtualized resource management costs and the routing cost.

The needs of computing cost of each BS-*k* to host its functions at the RU-*k*, vDU-*k* (in the FS) and vCU-*k* (in the ES) are denoted as:

$$f_{\mathrm{RU}}(\hat{w}_k^n), f_{\mathrm{FS}}(\hat{x}^n), \text{ and } f_{\mathrm{ES}}(\hat{y}_k^n), \quad (3)$$

where $f_{\mathrm{RU}}(\cdot)$, $f_{\mathrm{FS}}(\cdot)$ and $f_{\mathrm{ES}}(\cdot)$ are the cost functions to charge the utilized computing processing at the RU,[6] FS and ES, respectively. These cost functions translate the actual computing resource utilization of the RUs $\hat{w}^n \coloneqq \{\hat{w}_k^n \in \mathbb{R} : k \in \mathcal{K}\}$, vDUs $\hat{x}^n \coloneqq \{\hat{x}_k^n \in \mathbb{R} : k \in \mathcal{K}\}$ and vCUs $\hat{y}^n \coloneqq \{\hat{y}_k^n \in \mathbb{R} : k \in \mathcal{K}\}$ into monetary units ($). The actual resource utilization of each RU, vDU and vCU is highly affected by the split and demand at the BS. Hence, we define $\psi : (\lambda_k^n, i_k^n) \mapsto (\hat{w}_k^n, \hat{x}_k^n, \hat{y}_k^n)$ as a function to map inputs of the split and traffic demand of the BS into the actual resource utilization at the RU, vDU and vCU. This function represents the actual computing behavior in the vRAN system, and we characterize it through traces from the testbed measurements. Further, we consider that cost functions $f_{\mathrm{RU}}(\cdot)$, $f_{\mathrm{FS}}(\cdot)$ and $f_{\mathrm{ES}}(\cdot)$ to be proportional with their input, e.g., $f_{\mathrm{FS}}(v) \coloneqq \kappa_{\mathrm{RU}} v$, $f_{\mathrm{FS}}(v) \coloneqq \kappa_{\mathrm{FS}} v$ and $f_{\mathrm{ES}}(v) \coloneqq \kappa_{\mathrm{ES}} v$, where κ_{RU} ($/unit), κ_{FS} ($/unit) and κ_{ES}($/unit) are the estimated computing processing fees per core unit capacity at the RUs, FSs and ESs, respectively.

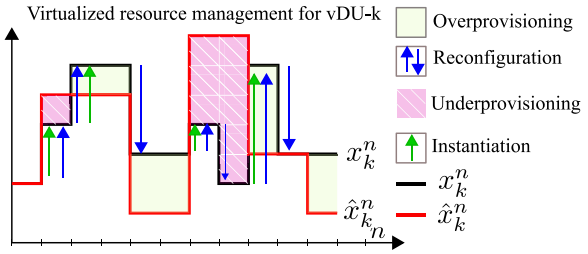


Fig. 4. An example of virtualized resource management model for vDU-*k*.

In vRANs, the vDUs and vCUs are virtualized on the FSs and ESs, respectively. Therefore, the virtualized resources of the vDUs and vCUs can be dynamically allocated to obtain cost-efficient network operations. However, reconfiguring such resources might lead to additional costs. Meanwhile, the allocated resources x_k^n and y_k^n might differ to the actual resource utilization of $\hat{x}_k^n$ and $\hat{y}_k^n$, which can create unwanted resource overprovisioning or declined demands. Motivated by resource management in network slicing [34], we propose a cost model capturing such behaviors in vRANs. This model is illustrated in Fig. 4 and described as follows.

(i) Overprovisioning: If the allocated resources are higher than their actual utilization, the operators pay more expenses and miss the opportunity to share their unused resources for other workloads. Such resources are instantiated and reserved for no purpose, which can be more profitable to be allocated for other workloads (e.g., video analytics) to increase the global system efficiency. This overprovisioning cost at time slot *n* for BS-*k* is defined as:

$$f_{\mathrm{O}}(\max(0, x_k^n - \hat{x}_k^n) + \max(0, y_k^n - \hat{y}_k^n)), \quad (4)$$

where $f_{\mathrm{O}}(\cdot)$ is a cost function for resource overprovisioning. This function is proportional with the input, e.g., $f_{\mathrm{O}}(v) \coloneqq \kappa_{\mathrm{O}} v$, where κ_{O} is the estimated fee for one unit capacity ($/unit).

(ii) Declined service demands: The declined demands can occur when there exists an insufficient resource allocation or constraint violation, which triggers service level agreement (SLA) violation and monetary compensation. For instance, the constraint violation can happen when the total allocated

[6]RUs are the radio hardware units; hence we do not allocate resources for RU. Instead, the computing cost of the RUs is incurred from processing the LP/RF functions, where their processing cost is demand/split dependent.

resources of the vDUs exceed FS capacity:

$$f_{\mathrm{D}}\left(\max\left(0, \sum_{k\in\mathcal{K}} x_k^n \mathbb{1}_{=l}(z_k^n) - H_l\right)\right),\ \forall l \in \mathcal{L}, \quad (5)$$

the total allocated resources of the vCUs exceed ES capacity:

$$f_{\mathrm{D}}\left(\max\left(0, \sum_{k\in\mathcal{K}} y_k^n \mathbb{1}_{=m}(\zeta_k^n) - \hat{H}_m\right)\right),\ \forall m \in \mathcal{M}, \quad (6)$$

and the incurred delay does not meet the requirement:

$$f_{\mathrm{D}}\Big(\max\Big(0, d_{p_{ml}} - d_i^H, d_{p_{lk}} - d_i^L, d_{p_{mk}} - 0.25\Big)\Big), \quad \forall m \in \mathcal{M}, \forall l \in \mathcal{L}, \quad (7)$$

where d_i^H and d_i^L are the delay requirement of split i for the HLS and LLS, respectively, as defined in Table I. In addition to the constraint violation, an insufficient allocation for each vDU and vCU can cause declined service demands, and we define this as:

$$f_{\mathrm{D}}(\max(0, \hat{x}_k^n - x_k^n, \hat{y}_k^n - y_k^n)). \quad (8)$$

The function $f_{\mathrm{D}}(\cdot)$ captures the monetary compensation that the operators have to pay for violating the SLA. This function is assumed to be proportional with the input, e.g., $f_{\mathrm{D}}(v) \coloneqq \kappa_{\mathrm{D}} v$, where κ_{D} is the estimated fee for declined demands in one unit capacity ($/unit).

(iii) Instantiation and Reconfiguration: The operators may decide to instantiate new resources or reconfigure their network settings to reduce resource overprovisioning and declined demands and adapt to the varying traffic demands and resource availability. However, instantiating and reconfiguring such resources (e.g., VMs) induce capital expenses, and we define it as:

$$f_{\mathrm{I}}\Big(\max\Big(0, x_k^n - x_k^{n-1}\Big) + \max\Big(0, y_k^n - y_k^{n-1}\Big)\Big), \quad (9)$$

$$f_{\mathrm{R}}\Big(\Big(|x_k^n - x_k^{n-1}| + |y_k^n - y_k^{n-1}|\Big) + \Big(x_k^n \mathbb{1}_{\neq z_k^{n-1}}(z_k^n) + y_k^n \mathbb{1}_{\neq \zeta_k^{n-1}}(\zeta_k^n)\Big)\Big), \quad (10)$$

where $f_{\mathrm{I}}(\cdot)$ and $f_{\mathrm{R}}(\cdot)$ are the cost functions for resource instantiation and reconfiguration. Eq. (9) captures the amount of instantiating additional resources for the vDU and vCU, which might arise due to migrating additional resources to serve the vRAN workload, and this results in indirect overhead expenses such as the increase of power consumption [34]. Then, the first term in (10) captures the reconfiguration cost initiated from migration activities for altering the splits and flavors (resizing resources). Such activities raise overhead costs from the migrated resources, measured from the difference between the current and the previous resources [9], [34]. For instance, altering the splits requires creating new BS functions while maintaining the old migrated functions to keep active [9]. Resizing the VMs' resources also initiates a price of management delay [35] as it needs time for migrating (and bootstrapping) the computing resources, load balancing and steering the network load.[7] The second term in (10) captures the reconfiguration cost for migrating the vDU and vCU instances to other FS and ES locations. In this case, the whole resources of vDU and vCU instances are affected, and the attached routing paths need to be recomputed with the new FS and ES locations. In our evaluation, $f_{\mathrm{I}}(\cdot)$ and $f_{\mathrm{R}}(\cdot)$ are proportional to the input, e.g., $f_{\mathrm{I}}(v) \coloneqq \kappa_{\mathrm{I}} v$ and $f_{\mathrm{R}}(v) \coloneqq \kappa_{\mathrm{R}} v$, where κ_{I} ($/unit) is the estimated cost for resource instantiation and κ_{R} ($/unit) is for reconfiguration. If reconfiguring the system does not incur any overhead cost, we can set $\kappa_{\mathrm{R}} = 0$, otherwise $\kappa_{\mathrm{R}} > 0$.

O-RAN has encouraged adopting an open interface between the vCUs, vDUs and RUs [5], resulting in sharing the xHaul links among the BSs. In addition, S1, S2, S3 and S4 generate different data loads depending on the selected split as seen in Table I. Hence, the cost for reserving bandwidth and routing the data flow through the xHaul links are also different. The routing cost for each BS-k can be denoted as:

$$f_{\mathrm{H}}\left(\sum_{p\in\mathcal{P}_k}\left(r_{p,i}^{\mathrm{FH},n}\sum_{l\in\mathcal{L}}\mathbb{1}_{=z_k^n}(l) + r_{p,i}^{\mathrm{MH},n}\sum_{m\in\mathcal{M}}\mathbb{1}_{=\zeta_k^n}(m) + r_{p,i}^{\mathrm{BH},n}\right)\right), \quad (11)$$

where $r_{p,i}^{\mathrm{FH},n}$, $r_{p,i}^{\mathrm{MH},n}$, $r_{p,i}^{\mathrm{BH},n}$ are the incurred data loads over FH, MH and BH at time slot n from using path p, serving traffic demand λ, and deploying split-i. The indicator $\mathbb{1}_{=z_k^n}(l)$ activates if vDU-k is placed at FS-l and $\mathbb{1}_{=\zeta_k^n}(m)$ activates if vCU-k is hosted at ES-m. Then, $f_{\mathrm{H}}(\cdot)$ is the cost function for bandwidth reservation to transfer data load through the xHaul links, and this cost function is proportional with the input, e.g., $f_{\mathrm{H}}(v) \coloneqq \kappa_{\mathrm{H}}^p v$, where κ_{H}^p ($/Gbps/Km) is the estimated fee for reserving bandwidth for path p per Gbps/Km.

Let suppose $J^n(a^n, s^n) \coloneqq \sum_{k\in\mathcal{K}} f_{\mathrm{RU}}(\cdot) + f_{\mathrm{FS}}(\cdot) + f_{\mathrm{ES}}(\cdot) + f_{\mathrm{O}}(\cdot) + f_{\mathrm{D}}(\cdot) + f_{\mathrm{I}}(\cdot) + f_{\mathrm{R}}(\cdot) + f_{\mathrm{H}}(\cdot)$ is the total operation cost for all the BSs accounted from (3)-(11). Then, we define the reward[8]:

$$r(a^n, s^n) \coloneqq -J^n(a^n, s^n). \quad (12)$$

Then, our aim is to design *an optimal policy* that maps the input state observation into action $\pi^*(s) : \mathcal{S} \mapsto \mathcal{A}$, which minimizes the long-term total operation cost over period of time. Such a policy can be formulated through maximizing the long-term reward:

$$\pi_* \coloneqq \arg\max \mathbb{E}_\pi\left[\sum_{\tau=0}^{N} \gamma^\tau r^{\tau+n} | \pi\right], \quad (13)$$

where $\mathbb{E}[\sum_{\tau=0}^{N} \gamma^\tau r^{\tau+n}]$ is the expected long-term accumulated reward starting at time slot τ. The discount factor γ is

[7]We have calculated the incurred time for resizing a VM instance in CSC cPouta (https://www.csc.fi/) cloud computing platform, and it takes around 25 seconds. Modern software architecture such as Kubernetes also requires several seconds to executing new pods [34].

[8]Our study focuses on network operation cost minimization, but our framework can be extended to other or multiple objectives, such as maximizing the vRAN performance (e.g., centralization degree). In this case, we can use weighting parameters that determine the relative importance between the objectives (e.g., cost and performance).

strictly set to $\gamma = 1$ during the online operation, corresponding to a non-discounted reward that represents the actual cost; otherwise, $\gamma \in (0, 1]$.

C. Trade-Offs

The above problem is intricate for many reasons. We discuss the trade-offs and non-triviality that arise as follow.

(i) From S1 to S4, the operators can gain a lower computational cost and high-performance operations through function centralization. However, it also has a tighter constraint requirement and induces a higher transferred data load through the xHaul links. A higher data load means a more expensive routing cost. In addition to the splits, the required resources for the vDUs and vCUs are highly affected by traffic demands and resource availability, which might change absurdly. These also affect the placement of the vDUs and vCUs over FSs and ESs. The association and routing paths are also different for each placement location.

(ii) Using a static policy and finding the best configurations by foreseeing the future peak traffic may reduce the overhead costs due to reconfiguration activities. However, it might produce significant resource overprovisioning. Such unused resources can be profitable if the operators can efficiently manage and share with other workloads. Predicting the future peak traffic might also be inaccurate, which might not result in the best configurations.

(iii) By dynamically reconfiguring the vRAN settings at every time slot, the operators can obtain the best configurations at a time; hence, the risks of resource overprovisioning and declined demands can be reduced. However, every reconfiguration activity produces overhead costs, which may lead to costly long-term network operations. Moreover, the reconfiguration decisions are made before the actual traffic demand is observed; therefore, finding the optimal decisions at every time slot is challenging and might be unfeasible in practice.

(iv) The reconfiguration decisions in our vRAN system are highly affected by the traffic demands and resource utilization. However, their relations are complex, depending on many factors such as traffic demand, computing platform, radio scheduler, etc, which also hinder general assumptions (e.g., linear) to model the computing resource's behavior, rendering traditional control policies inefficient for our vRAN reconfiguration problem.

(v) Points (i)-(iv) emphasize the need for intelligent reconfiguration decisions with minimal assumptions about the underlying system. A deep RL paradigm can be suitable to handle such challenges. However, the formulated RL problem has a huge state space and multi-dimensional action space because the vRAN system consists of multiple BSs sharing the same network resources with highly coupled configuration decisions. These challenges make conventional deep RL discrete action space algorithms such as deep Q learning inefficient.

Given the formulated RL problem and trade-offs above, we present how to design the solution that solves the problem efficiently in the next section.

IV. LARV Learning Algorithm

LARV leverages a model-free RL paradigm, which considers the vRAN system as a black-box environment and does not take any assumption about the system state and state transition probability distribution. However, finding the optimal policy of the agent is non-trivial as the formulated RL problem has the semi-continuous state space and the multi-dimensional action space, which make the state-action space extremely large. The large state space can be addressed using D3QN [15], where this approach is also naturally designed for discrete action. However, we need to tackle the issue of the multi-dimensional action space, which makes the number of estimated actions grow combinatorially with the number of BSs and configuration decisions. In order to address this curse dimensionality, we incorporate action branching [17] with D3QN to compress the number of estimated actions. Through this approach, the multi-dimensions of the action can be distributed across individual network branches while maintaining a shared decision module among them to encode a latent representation of the input state and enable coordination among the branches. In contrast to traditional discrete-action deep RL algorithms, this action decomposition method exhibits a linear growth of the total network outputs with increasing action dimensionality.

A. D3QN to Address the Large State Space

The objective of our RL agent is to learn the optimal policy π_* defined in (13). As the problem has a large state space and the expected output is a discrete action, we can utilize an off-policy RL algorithm by using D3QN to approximate the action-value function (Q-function) and Double Q-learning for the learning step.

We define the optimal action-value function $Q^*(s, a)$ as the maximum expected reward for observing certain sequences s after following some policies π and taking some actions a as: $Q^*(s, a) := \max_\pi \mathbb{E}[\sum_\tau^\infty \gamma r^{\tau+n} | s^n = s, a^n = a]$. If we know the optimal value $Q^*(s', a')$ of the sequence at the next time slot s' for all possible actions a', we can identify the optimal policy π^*, which is to select action a' that maximizes the expected value $r + \gamma Q^*(s', a') : Q^*(s, a) := \mathbb{E}_{s\sim\mathcal{E}}[r + \gamma \max_{a'} Q^*(s', a') | s', a']$. In the value iteration method, the action-value function can converge to the optimality when the iteration number reaches near infinity; however, it is impractical. Therefore, a function approximator such as a neural network can be applied to estimate the action-value function. The estimated action-value function parameterized by a neural network (Q-network) with weights θ is denoted as: $Q(s, a; \theta) \approx Q(s, a)$. Then, the Q-network is trained by minimization of a loss function:

$$L(\theta) := \mathbb{E}_{s,a,r,s'\sim\mathcal{D}}\big[u - Q(s, a; \theta)\big], \tag{14}$$

where the transition $\{s, a, r, s'\}$ is collected through random sampling (minibatches) from stored experience data $\mathcal{D}$, and u is the Temporal Difference (TD) target. In DQN [36], the TD target is computed by:

$$u^{\mathrm{DQN}} := \mathbb{E}_{s'\sim\mathcal{S}}\Big[r + \gamma \max \tilde{Q}\Big(s', a'; \tilde{\theta}\Big)\Big], \tag{15}$$

where $\tilde{Q}(s', a'; \tilde{\theta})$ is the target network parameterized by weights $\tilde{\theta}$. The design of TD-target in (15) often causes an overestimate to the actual action-value. Thus, we apply Double DQN (DDQN) [16] to overcome this issue by modifying the TD target into:

$$u^{\text{DDQN}} \coloneqq \mathbb{E}_{s' \sim \mathcal{S}}\left[r + \gamma \tilde{Q}\left(s', \arg\max_{a'} Q\left(s', a'; \theta\right); \tilde{\theta}\right)\right]. \tag{16}$$

When the RL problem has a large action space, such as in our vRAN problem, it might not require estimating the value for certain states, i.e., avoiding unnecessary estimation of redundant and low-value actions. Thus, we apply the Dueling architecture [15] to DDQN (called D3QN) by separating the Q-network into two streams of state-value and advantage, which are then combined through an aggregating layer to produce an estimate of the action-value function. Lets denote $V(s;\theta)$ and $A(s, a;\theta)$ as the estimated state-value function and advantage function, respectively; then, the action-value function at the output layer can be computed as:

$$Q(s, a;\theta) \coloneqq V(s;\theta) + A(s, a;\theta) - \frac{1}{|\mathcal{A}|}\sum_{a'} A\left(s, a';\theta\right). \tag{17}$$

By explicitly separating the Q network into two estimators, D3QN can learn which states are valuable without requiring to learn the impact of every action for each state. Hence, it can effectively achieve a high-quality policy for a large state space. However, in addition to a large state space, our vRAN problem produces a multi-dimensional discrete action space. It drives the number of estimated Q values in (17) to grow combinatorially with the number of configuration decisions and BSs. Next, we present how we incorporate an action branching architecture with D3QN to compress the number of estimated Q values in our vRAN problem.

B. Action Compression Using Action Branching

Let us define $\mathcal{C}_k \coloneqq \{i_k, x_k, y_k, z_k, \zeta_k\}$ as a set that includes all the reconfiguration control variables of BS-k. Then, we denote the sub-action $a_{kc}, \forall c \in \mathcal{C}_k, \forall k \in \mathcal{K}$, to represent the c-th reconfiguration control variables of BS-k, i.e., $a_{11} \coloneqq i_1, a_{12} \coloneqq x_1, \ldots, a_{KC_K} \coloneqq \zeta_K$; and $C_k \coloneqq |\mathcal{C}_k|, \forall k \in \mathcal{K}$. Hence, we can rewrite the action in (1) by $a \coloneqq \{a_{kc} : c \in \mathcal{C}_k, k \in \mathcal{K}\}$. Each of sub-actions also takes values from a finite set of the sub-action space $\mathcal{A}_{kc} \subseteq \mathcal{A}$ that describes the c-th reconfiguration control space of BS-k, i.e., $\mathcal{A}_{k1} \coloneqq \mathcal{I}, \mathcal{A}_{k2} \coloneqq \mathcal{X}, \ldots, \mathcal{A}_{kC_K} \coloneqq \mathcal{M}, \forall k \in \mathcal{K}$. As the RL agent controls K BSs, and each BS has C_k sub-actions; then, the number of Q-values to be estimated turn to $\prod_{k=1}^{K}\prod_{c=1}^{C_k}|\mathcal{A}_{kc}|$. By incorporating action branching, the number of Q-values to be estimated can be compressed to $\sum_{k=1}^{K}\sum_{c=1}^{C_k}|\mathcal{A}_{kc}|$. The initial action branching in [17] has successfully tackled problems with the discretized continuous action space. However, its performance is still not validated in the problem where the action space is naturally multi-dimensional. Moreover, it assumes that all of the sub-action spaces have the same dimensional size, i.e., $|\mathcal{A}_{11}| = |\mathcal{A}_{12}| = \cdots = |\mathcal{A}_{KC_K}|$. Hence, we can not directly utilize it as the size of the sub-action space of the reconfiguration control variables in our vRAN problem varies. We adopt the action branching paradigm suited to our problem and describe it as follows.

We use the common state s defined in (2) and common state-value $V(s)$. The value of sub-action a_{kc} at common state s with the corresponding sub-action advantage $A_{kc}(s, a_{kc})$ becomes:

$$Q_{kc}(s, a_{kc}) \coloneqq V(s) + \left(A_{kc}(s, a_{kc}) - \frac{1}{|\mathcal{A}_{kc}|}\sum_{a'_{kc} \in \mathcal{A}_{kc}} A_{kc}\left(s, a'_{kc}\right)\right). \tag{18}$$

Then, the TD target is set similar to (16) to avoid maximization bias, except it uses an average of all the dimensions of the sub-actions as follows:

$$u \coloneqq r + \gamma \frac{1}{K}\sum_{k=1}^{K}\frac{1}{C_k}\sum_{c=1}^{C_k}\tilde{Q}_{kc}\left(s', \arg\max_{a_{kc'} \in \mathcal{A}_{kc}} Q_{kc}\left(s', a'_{kc}\right)\right), \tag{19}$$

where $\tilde{Q}_{kc}$ is the target network. Then, the loss function can be computed as:

$$L(\theta) \coloneqq \mathbb{E}_{s,a,r,s' \sim \mathcal{D}}\left[\frac{1}{K}\sum_{k=1}^{K}\frac{1}{C_k}\sum_{c=1}^{C_k}[u_{kc} - Q_{kc}(s, a_{kc};\theta)]\right]. \tag{20}$$

The action a to be taken for all the BSs is selected based on ϵ-greedy, where the agent chooses a random action with probability ϵ or compute:

$$a \coloneqq \left[\arg\max_{a'_{k1}} Q_{k1}(s, a_{k1'}), \ldots, \arg\max_{a'_{KC}} Q_{KC_K}\left(s, a_{KC'_K}\right)\right] \tag{21}$$

with probability $1 - \epsilon$.

C. Neural Network Architecture and Learning Algorithm

Fig. 5 illustrates the Q-network architecture of branching D3QN Q_θ, parameterized by weights θ and applied in LARV. This network is constructed from an input layer, a shared representation segment comprising hidden layers, a state value network, and neural network branches. The input layer (Linear layer with ReLU activation) receives the common state observation s and has the size of $|s|$. The shared representation segment is built from two fully connected Linear layers with ReLU activation, connected to neural network branches and state value function network. We use a Linear layer for the common state value network. Then, the neural network branches have a total of $\sum_{k=1}^{K} C_k$ branches corresponding to the number of control decision variables (sub-actions). Each branch aims to produce the sub-action value $Q_{kc}(s, a_{kc})$ by taking consideration of the common state value $V(s)$ and sub-action advantages $A_{kc}(s, a_{kc})$ as described in (18). Each

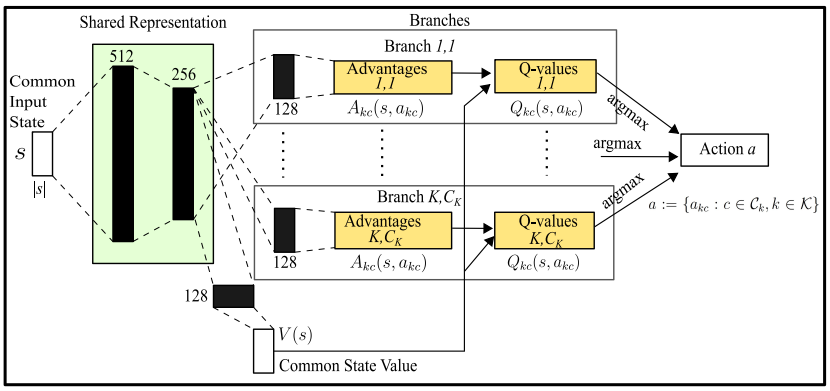


Fig. 5. The Q-network architecture built following D3QN with action branching.

Algorithm 1: LARV Learning Algorithm

Initialize: Replay memory $\mathcal{D}$ with a fixed buffer size, Q-network Q_θ (Fig. 5) with random or pretraining weights θ.
Clone Q-network Q_θ to target network $\tilde{Q}_{\tilde{\theta}}$ with weights $\tilde{\theta} \leftarrow \theta$.
for *Each episode* $e = 1.., E$ **do**
  Reset state of all the BSs $s^1 := \{\lambda^1, i^0, x^0, y^0, z^0, \zeta^0\}$.
  for *Each time slot* $n = 1..., N$ **do**
    Select an action $a^n := \{a^n_{kc} : c \in \mathcal{C}_k, k \in \mathcal{K}\}$ randomly with probability ϵ, otherwise compute a^n by using (21).
    Determine the routing $p \in \mathcal{P}_k, \forall k \in \mathcal{K}$ using i^n, z^n and ζ^n obtained from a^n.
    Enforce a^n and $p \in \mathcal{P}_k, \forall k \in \mathcal{K}$ to all the BSs and compute the total cost J^n.
    Collect the reward r^n based on (12).
    Set $s^{n+1} \leftarrow s^n$ with the current observation.
    Store the experience $\mathcal{D} \leftarrow \{s^n, a^n, r^n, s^{n+1}\}$.
    Sample minibatch of experiences from $\mathcal{D}$.
    Compute TD target u using (19) if not done, otherwise $u := r^n$.
    Perform a gradient descent method to the loss function $L(\theta)$ in (20) w.r.t θ.
    Update target network $\tilde{Q}_{\tilde{\theta}} \leftarrow Q_\theta$ every $\hat{n}$ steps.
  end
end

branch has an output layer (an aggregation layer from the state value and sub-action advantages) with the size of $|\mathcal{A}_{kc}|$.

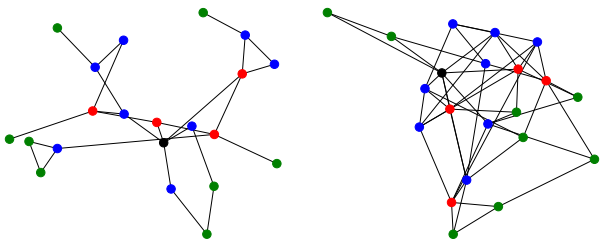
Further, we summarize the learning process of LARV in Algorithm 1. Firstly, the replay buffer memory $\mathcal{D}$ and the Q-network Q_θ (Fig. 5) are initialized, where the Q-network initialization can be from random or pretrained weights **(Step 1)**. Then, the weights of the Q-network Q_θ are copied to the target network $\tilde{Q}_{\tilde{\theta}}$ **(Step 2)**. At the beginning of each episode (or trial during the training), the state observation s^1 is reset with initial values, where these values are assigned from $\lambda^1 := \{\lambda^1_k \in \mathbb{R}_+ : k \in \mathcal{K}\}$, $i^0 := \{i^0_k = \text{S1} : k \in \mathcal{K}\}$, $x^0 := \{x^0_k = \max(\mathcal{X}) : k \in \mathcal{K}\}$, $y^0 := \{y^0_k = \max(\mathcal{X}) : k \in \mathcal{K}\}$, $z^0 := \{z^0_k = \text{random}(\mathcal{X}) : k \in \mathcal{K}\}$ and $\zeta^0 := \{\zeta^0_k = \text{random}(\mathcal{X}) : k \in \mathcal{K}\}$ **(Step 4)**. Then, at every time slot n, given the state observation s^n, an action $a^n := \{a^n_{kc} : c \in \mathcal{C}_k, k \in \mathcal{K}\}$ is selected randomly with probability ϵ, otherwise it is computed using (21) **(Step 6)**. Then, the routing $p := p_{0m} \cup p_{ml} \cup p_{lk} \in \mathcal{P}_k : k \in \mathcal{K}$ can be selected through i^n, z^n, and ζ^n obtained from the selected action since these variables determine the hosting servers for the vDUs and vCUs, and hence the destination server for each data flow **(Step 7)**. After all the control variables are determined, they are enforced to all the BSs as the vRAN configurations at time slot n. As a result of the deployed configurations, LARV expects to receive the total operation cost $J(a^n, s^n)$ **(Step 8)**. Based on this cost, the reward $r(a^n, s^n)$ signal at time n can be computed by (12) **(Step 9)**. The state is updated with the current observation $s^{n+1} \leftarrow s^n$ **(Step 10)**. Then, the agent's experience is stored in replay memory $\mathcal{D} \leftarrow \{s^n, a^n, r^n, s^{n+1}\}$ **(Step 11)** and the memory $\mathcal{D}$ is sampled randomly **(Step 12)**. Further, the TD target of branching D3QN is computed with (19). Once the TD-target is obtained, we can proceed to calculate the loss function $L(\theta)$ using (20) **(Step 13)**. The goal of this learning process is to minimize this loss function with regards to weights θ, and we rely on Adam optimizer [37] to perform stochastic gradient descent. Mostly, the target network is frozen, but it is updated every $\hat{n}$ by using the Q-network weights **(Step 15)**.

V. Results and Discussion

In this section, we perform trace-driven simulations using real traces collected from our testbed to evaluate the performance of LARV under various scenarios during the training process and online operation.

A. Experimental Setup

We built a bespoke testbed to collect measurements used to evaluate LARV under realistic conditions. We utilize the software-based srsRAN [12], where each entity is virtualized using container-based virtualization from Docker. The radio interfaces of the BS (e.g., RU) and user are emulated via ZMQ. The srsENB acts as a BBU of the BS. To deal with functional split, we use prior studies that divide the computing



(a) (b)

Fig. 6. The network graph representation for (a) N1 and (b) N2. The green, blue, red, and black dots represent the RUs, FSs, ESs and EPC, respectively.

consumptions of LP, HP, LM, HM, LR, HR, and PD functions to yield 48%, 17%, 7%, 7%, 0.5%, 0.5%, 10%, 10% of the total BBU, respectively, cf. [18], [20]. We deploy the virtualized entities in Platform A (CSC cPouta hpc.5.16core with max. 16 vCPU) and Platform B (PC AMD Ryzen 7 PRO 4750U with max. 16 CPU threads). We use these computing specifications for *Reference Core (RC)*, i.e., 1 RC translates to 1 CPU thread and 1 vCPU. The virtualized resource of each container can be controlled through *–cpus*, which allows us to set a capacity limit and isolate each container resource. We set an initial resource reservation for srsENB with 10 RCs. In our measurements, the traffic demand follows a Poisson-generated user datagram protocol with a peak data rate is 36.6 Mbps (SISO 10 MHz LTE).

In our simulations, the traffic demands follow the Milan network datasets from Telecom Italia [38], where each time slot has 10 minutes time interval. This interval is also aligned with the capabilities of current Virtual Infrastructure Managers (VIMs). Moreover, LARV selects an action from the incoming state information (e.g., by passing forward through the Q network) at each time slot, and it can be performed within a second in our test, which is suitable for real-time operation. The Milan datasets consist of mixed traffic, including calls, sms, and the Internet. We filtered the datasets and utilized Internet traffic (mobile broadband). Although it was recorded in 2013 (dominated by 4G traffic), it is still relevant for 5G network evaluation since it captures users' demand behavior comprehensively (e.g., the day, night, weekend, city center, etc.). Considering the limitations of our testbed and the difficulty in capturing the computing behavior of the Milan traffic in a tractable model, we utilize a deep neural network[9] to map the Milan traffic demands into the actual resource utilization, trained using our collected measurements.

We consider a realistic MEC-based Milan topology (N1) [39] and a synthetic topology (N2) generated using the Waxman algorithm [40], and their graph representation is illustrated in Fig. 6. N2 has parameters of link probability (0.5) and edge length control (0.1). A vRAN system in N1 and N2 consists of 1 EPC, 4 ESs, 8 FSs and 8 RUs (default), where the routers are co-located with each node.[10] Per link's latency, capacity, and weights of N1 and N2 vary from 0 to 0.1 ms, 30 Gbps to 160 Gbps, and 0 to 0.1. We have $H_l = 20\text{RCs}, \forall l \in \mathcal{L}$ and $\hat{H}_m = 100\text{RCs}, \forall m \in \mathcal{M}$. We set the available flavors with $|\mathcal{X}| = 16$ for Platform A and Platform B, which translate to $\{0, 1, \ldots, 14, 15\}$ RCs of the computing resources. Then, we define two vRAN systems in which we utilize Platform A with N1 (VR1) and Platform B with N2 (VR2).

We set the computing processing fee (per CPU usage) at the RU with $\kappa_{\text{RU}} \coloneqq 1\text{RC}^{-1}$ [18]. A single ES can serve up to 8 FSs, and a single FS can handle as high as 8 RUs. Therefore, we set $\kappa_{\text{FS}} \coloneqq 0.5\kappa_{\text{RU}}$ and $\kappa_{\text{ES}} \coloneqq 0.5\kappa_{\text{FS}}$ (c.f. [41, Fig. 6a] with $\approx$ 10 BSs) by taking into account the processing gain from centralization (i.e., computational processing cost is less by centralizing more functions and executing them in a higher computing platform). Then, with regards to prior study in [34], we set the coefficient fee for resource overprovisioning with $\kappa_{\text{O}} \coloneqq 1\text{RC}^{-1}$ and declined demands with $\kappa_{\text{D}} \coloneqq 5\text{RC}^{-1}$. It is common that the penalty due to the declined demands incurs a higher cost. We also set the default coefficient for the reconfiguration fee lower with $\kappa_{\text{R}} = 0.1\text{RC}^{-1}$ to account for the typically relatively lower cost per unit of resource reconfiguration [34]. Then, we set $\kappa_{\text{I}} \coloneqq \kappa_{\text{R}}$ (see Section III) and $\kappa_{\text{H}} \coloneqq 1\text{Gbps}^{-1}/\text{Km}$ (e.g., the fee for reserving 1 Gbps/Km routing bandwidth is the same as a processing fee at RU).

The Q-network of branching D3QN has an input layer with size of $|s|$, hidden layers (the architecture and size are provided in Fig. 5), and $\sum_{k=1}^{K} C_k$ branches. Each branch has an output with size of $|\mathcal{A}_{kc}|$. The target network is updated every 500 time slots. The batch size is set with 128 and the replay buffer has a capacity of 10^6. Our exploration and exploitation strategy is based on ϵ-greedy, where we set $\epsilon_{\max} = 1$ at the beginning of episode, then it exponentially decays to $\epsilon_{\min} = 0.015$. We use Adam optimizer [37] with learning rate is set to 0.0001 and (20) for the loss function. The time horizon for a single episodic training is one day ($N = 144$ time slots) and the online operation starts on the second day with a default duration of two days ($N = 288$ time slots). Table III summarizes the default experimental setups used in our evaluation. The datasets in this work will be released online.[11]

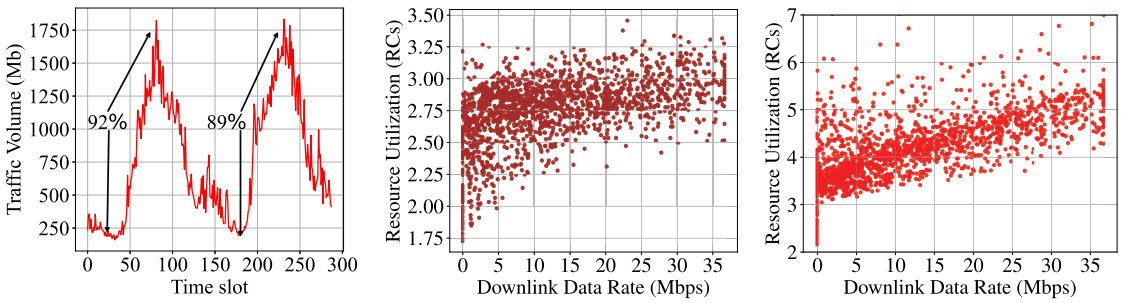
Further, we compare LARV with several benchmarks as follows.

- *The best static with 100% provisioning (BSP):* It knows exactly the peak future traffic demand of each BS and utilizes them to find the best static joint action via an exhaustive search. It can be defined as: $\pi_{BSP} \coloneqq \arg\min_a \sum_k^K J_k^i(a)$, where $i = \arg\max_n \lambda_k^n$. Further, it is used to normalized the monetary costs in the online operation evaluations.
- *DDPG with discretization:* Since the state space and action space of the RL problem are extremely large, the traditional discrete RL algorithm may not perform efficiently. A continuous RL algorithms such as DDPG [42]

[9]It is constructed from an input, an output and three hidden layers with the sizes of 128, 64 and 16. We use Adam optimizer [37] with learning rate is set to 5×10^{-5}, mini-batch with the size of 128 and MSE loss function, then train it with 200 epochs.

[10]The datasets for N1 and N2 initially do not specify which nodes are for EPC, ESs, FSs, and RUs. We followed an intuitive approach by selecting them from the highest network degree.

[11]https://github.com/fahriwm/larv_datasets



(a) (b) (c)

Fig. 7. **a)** Traffic variation within two days from Milan datasets [38] and **b)** collected measurement results over Platform A and **c)** Platform B. The resource utilization is presented in a reference core (RC), which translates to 1 virtual CPU/thread.

TABLE III
EXPERIMENTAL SETUP; SEE SECTION V-A FOR DESCRIPTION

Parameters	**Default value**
Number of ESs (M)	4
Number of FSs (L)	8
Number of RUs (K)	8
FS computing capacity (H_l)	20 RCs
ES computing capacity ($\hat{H}_m$)	100 RCs
The set of flavors ($\mathcal{X}$)	$\{0, 1, 2, ..., 15\}$ RCs
Overprovisioning fee (κ_{O})	1 RC^{-1}
Declined demand fee (κ_{D})	5 RC^{-1}
Reconfiguration fee (κ_{R})	0.1 RC^{-1}
Instantiation fee (κ_{I})	0.1 RC^{-1}
Processing fee at ES (κ_{ES})	0.25 RC^{-1}
Processing fee at FS (κ_{FS})	0.5 RC^{-1}
Processing fee at RU (κ_{RU})	1 RC^{-1}
Bandwidth (routing) fee (κ_{H})	1 Gbps^{-1}/Km
Time horizon (N) for 1 episode	144 time slots
Epsilon start and end ($\epsilon_{\max}, \epsilon_{\min}$)	(1, 0.015)
Learning rate	0.0001
Batch size	128

can address extremely large state-action space, but they are not designed for a discrete action. Hence, we relax the discrete action (1) into a continuous action. Then, when the output of DDPG is determined, we estimate it to the nearest discrete value. We also modify the output activation function with a Sigmoid function as each action needs to be a positive value.

- *Multi-agent of D3QN (MDQ):* It is a non-branching D3QN. To deal with multi-dimensional action space, in every BS, each reconfiguration control works as a separate agent, i.e., the decision of each split, resource, and location is controlled by a different agent, that works collaboratively to maximize the common reward in (12). In total, MDQ has $\sum_{k=1}^{K} C_k$ agents. The agents that represent control variables in the same BS share a common state observation.

B. Measurement Insight

Fig. 7(a) illustrates an example of the traffic demand of a BS in the Milan datasets [38]. It shows a significant difference between the peak and lowest traffic demand by up to 92% in a single day. Moreover, the traffic variation might vary from day to day (e.g., weekdays, weekends). Figs. 7(b) and 7(c) show that the traffic demand highly affects the resource utilization of the BBU. These findings motivate us to implement the dynamic configurations to adapt such traffic and resource variations to achieve cost-effective operations. Figs. 7(b) and 7(c) also demonstrate that the relations between traffic demand and resource utilization have high variance, where we found a significant degree of spread on the resource utilization. Moreover, these relations are platform-dependent performance (e.g., hanging on the hosting platforms and platform load). For example, although they indicate not strongly linear in Platform A and B, the resulting Pearson coefficient is different with 0.513 and 0.654, respectively. Also, albeit the BBU has been reserved with the same resources, Fig. 7(c) shows that the BBU utilization of Platform B is higher than Platform A. Such platform-dependent performance is also found in [10] for uplink, where the computing behavior of vRANs is identified depending on many latent factors.

C. Performance During Training Process

1) Training Convergence: Fig. 8 illustrates the convergence behavior of LARV over various reconfiguration coefficient fees in VR1 and VR2. At the beginning of episodes, LARV has a higher probability of utilizing a random policy for exploration. As a result, LARV produces a high long-term total operation cost over all the reconfiguration fees in VR1 and VR2. However, after some episodes, LARV successfully learns the optimal policy, starts to act greedily with a high probability, and convergences to the best policy the agent can learn. Moreover, we found a similar trend in LARV's behavior, where it manages to converge to some cost values after 400 episodes, albeit it learns over different reconfiguration fees and vRAN systems.

Fig. 8 also shows that using a random policy in vRAN reconfiguration problem must be avoided as it yields in costly long-term cost. In VR1, our findings reveal that LARV can save the costs by up to 78.14%, 79.0%, 80.76% and 83.2% over $\kappa_{\mathrm{R}} = 0.05$, $\kappa_{\mathrm{R}} = 0.1$, $\kappa_{\mathrm{R}} = 0.5$ and $\kappa_{\mathrm{R}} = 1$, respectively, compared to a random policy. Such significant cost savings by LARV also appear in VR2, where LARV can save the cost as high as 75.79%. The cost savings of LARV also

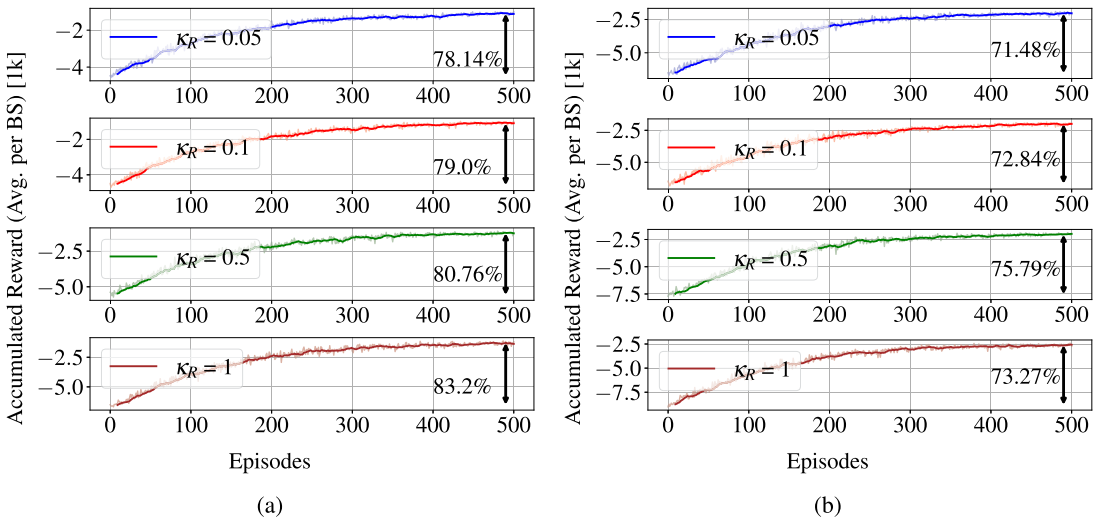


Fig. 8. The convergence of LARV under various reconfiguration fees in **a)** VR1 and **b)** VR2.

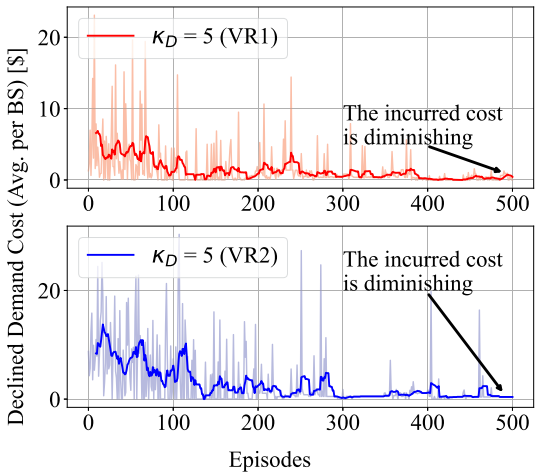


Fig. 9. The incurred cost from declined service demands (average per BS) in VR1 and VR2. The cost is diminishing as the training goes, and it eventually reaches to near zero (e.g., after 400 episodes).

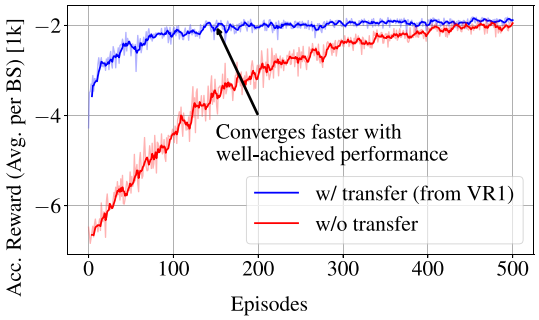


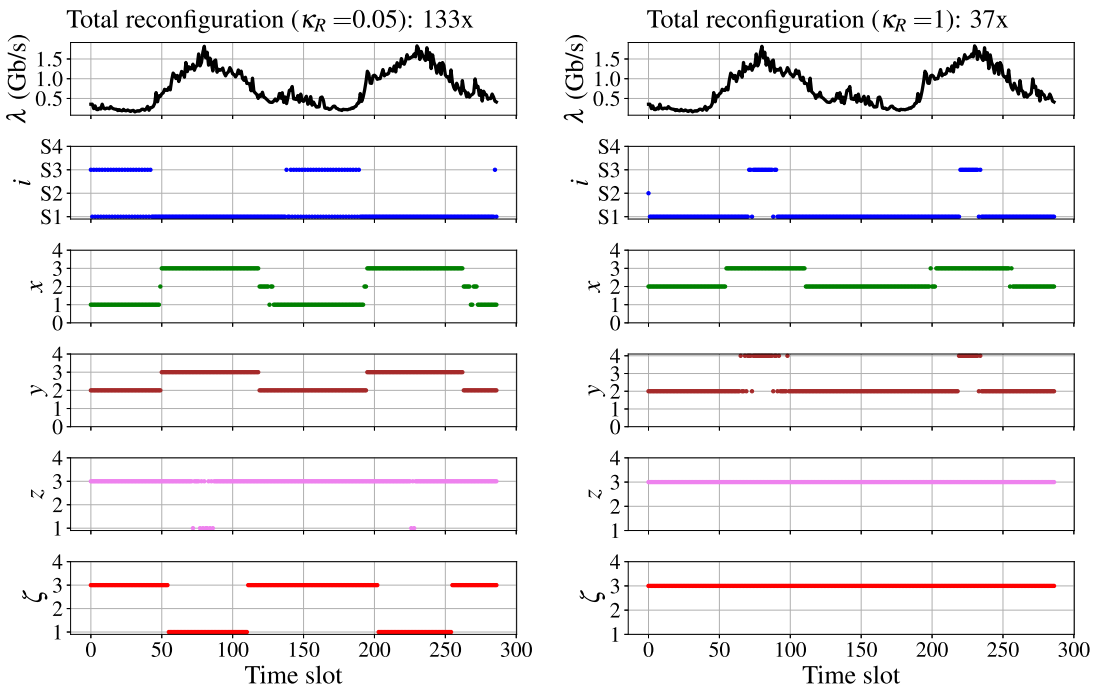
Fig. 10. Training convergence in VR2. Using transfer learning paradigm ("w/ transfer"), which is leveraged from pretraining weights in VR1, can achieve similar performance and faster convergence compared to without transfer learning ("w/o transfer").

increase when the reconfiguration fee is more expensive (e.g., $\kappa_{\mathrm{R}} = 0.05$ to $\kappa_{\mathrm{R}} = 1$).

2) Declined Demands: Fig. 9 shows that LARV can reduce the incurred cost due to declined demands after several training episodes both in VR1 and VR2. The declined demand cost appears in almost every episode at the beginning of training episodes. The main reason is that LARV mostly chooses random actions for exploration, rendering a very high number of declined service demands and, at the same time, producing a very expensive cost. Note that the declined demands contribute a significantly more expensive cost as its coefficient fee is much higher than others. As the training continues, LARV optimizes its weights based on the reward feedback and successfully diminishes the declined demand cost. After around 400 episodes, the incurred cost at each episode becomes smaller and less frequent, eventually reaching almost zero (or zero). Following the decrease of this cost, at the same time, the accumulated total operation cost (see Fig. 8) is also greatly diminished.

3) Transfer Learning: To assess the generalization of LARV over heterogeneous vRAN systems, we study the benefits of utilizing a transfer learning paradigm ("w/ transfer") compared to learning from scratch ("w/o transfer"). In particular, we leverage our pre-trained neural network weights (trained in VR1) for initializing the other neural network weights in different vRAN systems (e.g., in VR2). It is worth noting that the system parameters and platforms in VR1 and VR2 are different. Hence, this evaluation aims to study the possibility of reusing the existing models for the other vRAN systems, which might expedite the convergence and widespread deployment of LARV. We use the same default hyperparameter (defined in Section V-A), except we encourage less exploration for "w/ transfer" by modifying $\epsilon_{\max} = 1$ to $\epsilon_{\max} = 0.1$.

Fig. 10 depicts that LARV "w/ transfer" successfully converges to the similar value with "w/o transfer" in VR2, albeit the pre-trained weights are leveraged from a different vRAN system (VR1). Moreover, "w/ transfer" can speed up the training convergence with similar performance as "w/o transfer" even though the pre-training is conducted not in the same platform, where it starts to converge after around 150 episodes.



(a) $\kappa_{\mathrm{R}} = 0.05$ (b) $\kappa_{\mathrm{R}} = 1$

Fig. 11. The selected actions over different reconfiguration fees and traffic demand variations at BS-1 during online operation (2 days).

In transfer learning, a pre-trained model is utilized. And when a pre-trained model is available, the gained knowledge of this already trained model can be transferred among *different but similar (e.g., correlated)* environments and contexts, which in our case are VR1 and VR2. Such a transfer knowledge paradigm can expedite the learning convergence and allow the reuse of existing pre-trained models across different but related vRAN systems (i.e., have correlations with the training environment/context).

4) Action Space Compression: Following the simulation setup, each BS has sub-action sizes with $|\mathcal{I}| = 4$, $|\mathcal{X}| = 16$, $|\mathcal{L}| = 8$ and $|\mathcal{M}| = 4$, and we have $|\mathcal{K}| = 8$. Hence, the number of Q values to be estimated is originally around 1.32×10^{36}. LARV turns such a combinatorial explosion into a linear increase; hence, the number of estimated Q values becomes 384.

D. Performance During Online Operation

1) Selected Actions: Fig. 11 illustrates how LARV successfully controls the configurations of BS-1 reacting to the traffic variations and resource availability over different reconfiguration fees. Instead of minimizing the incurred cost at each slot, LARV's objective is to minimize the cost in the long run. As shown in Fig. 11(b), LARV performs 133× reconfiguration activities when $\kappa_{\mathrm{R}} = 0.05$. However, this activity becomes less frequent with the increase of reconfiguration fee, where there are only 33× reconfiguration activities. For the functional split (i), LARV mostly selects S1 (more decentralized functions) when the traffic of BS-1 is low, and it adjusts the split decision to S3 (more centralized functions) as the traffic increases. In S3, the transferred data flow over HLS is equal to the traffic demand with 500 Mbps of additional signaling overhead ($\lambda + 0.5$ Gbps); hence, LARV does not suggest implementing it in low traffic for such high overhead. However, when the traffic is elevated, i.e., the signaling does not significantly contribute to the data flow and routing cost, LARV tends to choose S3, considering the benefits of function centralization. This behavior appears in both $\kappa_{\mathrm{R}} = 0.05$ and $\kappa_{\mathrm{R}} = 1$, though the number of reconfigurations differs, where the reconfiguration is more often for $\kappa_{\mathrm{R}} = 0.05$. Further, the other results suggest that the allocated resources and the placement locations for vDUs and vCUs vary for different reconfiguration fees. For instance, the allocated resource of the vDU (x) in $\kappa_{\mathrm{R}} = 1$ is larger than in $\kappa_{\mathrm{R}} = 0.05$, even during the traffic is low, as LARV needs to accommodate the less frequent reconfigurations and more decentralized functions (it mostly implements S1). LARV also directly allocates a higher resource of the vCU (y) to avoid numerous reconfigurations when $\kappa_{\mathrm{R}} = 1$. Moreover, LARV decides to rarely reconfigure the vDU (z) and vCU (ζ) locations or even does not reconfigure them when the fee is costly ($\kappa_{\mathrm{R}} = 1$), as altering such configurations requires migrating all the resources to the new places, which can trigger significantly expensive reconfiguration cost.

2) The Number of BSs: We evaluate LARV over a different number of the BSs in the vRAN system and present it in Fig. 12(a). The number of BSs significantly influences the size of the state space and action space of the RL problem. In general, all the RL approaches outperform BSP when $K = 1$, where LARV becomes the most cost-effective by saving the cost up to 59%. However, when the number of BSs in the vRAN system becomes more prominent, the size of the action

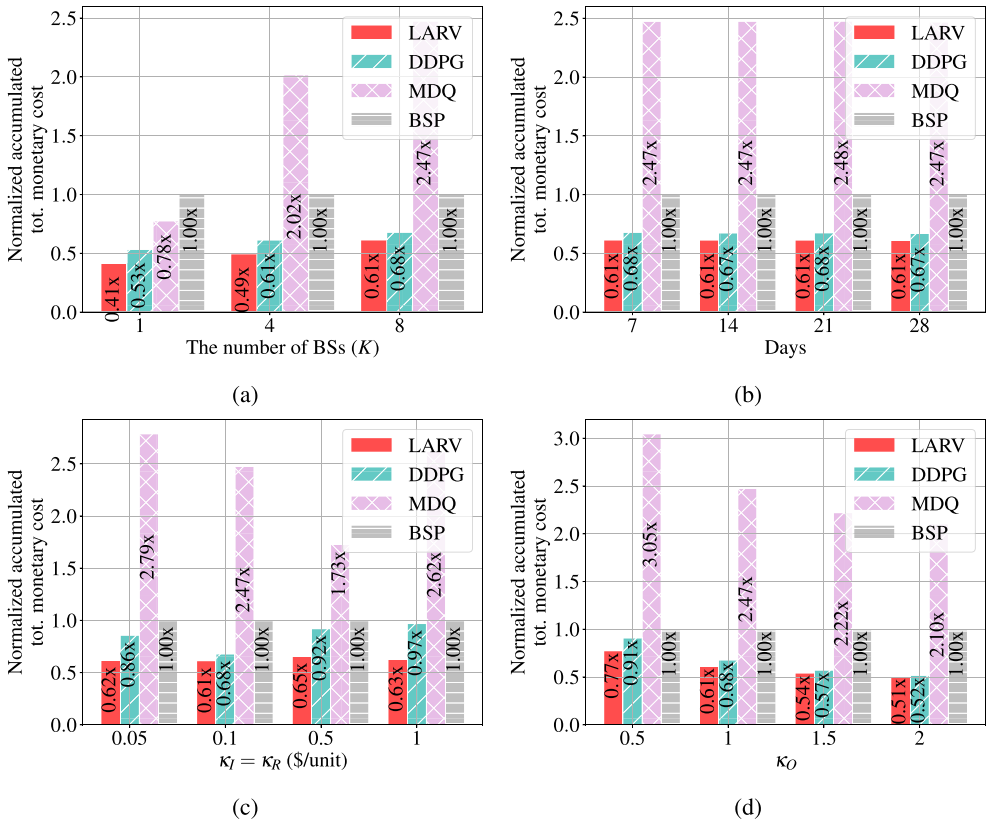


Fig. 12. Performance during the online operation in VR1. The presented monetary costs are normalized to BSP.

space, state space, and the number of possible actions grow combinatorially. By adopting action branching, LARV successfully deals with such a combinatorial growth with a linear increase, rendering well-achieved performance, as shown in Fig. 12(a). And it brings LARV to be the least degraded performance, where the cost savings of LARV is more than 39% of BSP. In contrast to LARV, MDQ utilizes a distributed multi-agent system. When the number of BSs increases, the number of agents of MDQ also increases, and this makes the performance of MDQ deteriorate compared to the centralized learning approaches. Moreover, albeit DDPG can deal with discrete action space through discretization of continuous action, the performance is still far from LARV. Unlike LARV, which is naturally designed for large discrete action space, DDPG can lose its learning effectiveness due to discretization.

3) Time Horizon Setting: Fig. 12(b) visualizes the performance of LARV compared to the benchmarks over various time horizon settings, ranging from 7 days to 28 days. We found that LARV becomes the most cost-effective approach by having the cheapest long-term total cost. The performance of LARV also remains stable, albeit in varying conditions (demands and resource availability). Compared to BSP, the cost-savings of LARV can be as high as 39%. LARV updates the vRAN configurations prudently, adapting to the varying conditions and considering the long-term cost, while BSP follows static policy by using future traffic information. This finding clearly emphasizes the importance of dynamic reconfiguration in vRANs. Moreover, LARV also outperforms RL benchmarks, where it saves the long-term total cost by up to 10% of DDPG and 75% MDQ. Compared to continuous space and non-branching state-of-the-art deep RL approaches, this gain shows the effectiveness of LARV through branching of D3QN in solving a large state space and multi-dimensional action space of the vRAN reconfiguration problem.

4) Reconfiguration Fees: We analyze the impact of various reconfiguration fees (κ_{R}) on the cost savings that LARV can achieve. Fig. 12(c) shows that LARV can successfully provide well-achieve performance in both cheap and expensive reconfiguration fees. It also shows that the increase in reconfiguration fee slightly affects the performance of LARV while it significantly degrades DDPG. DDPG is proposed for continuous action, and the performance can be deteriorated due to discretization when the problem has discrete action space, such as arising in our problem. In general, compared to DDPG, the cost savings of LARV increase as the fee gets more expensive, where the gains of LARV rise from 10% to as high as 35% ($\kappa_{\mathrm{R}} = 1$). Moreover, the cost savings of LARV remain stable compared to BSP and MDQ at around 35-39% and 62-76%, respectively. These findings emphasize that reconfiguring the vRAN system is beneficial, but we need to carefully design the RL algorithm suited to the vRAN problem.

5) Overprovisioning Fees: We study the effect of different overprovisioning coefficient fees to the performance of LARV. As seen from Fig. 12(d), when the overprovisioning fee gets costly, all the RL approaches' performance increases correspond to the static policy, where LARV becomes the best approach among them. LARV can save the costs from 23% ($\kappa_{\mathrm{O}} = 0.05$) to 49% compared to BSP ($\kappa_{\mathrm{O}} = 1$).

These results highlight the need for reconfiguring prudently the vRAN system at runtime, particularly when the resources are valuable and the price of wasting such resources is high, making the static policy economically unviable for long-term operations.

VI. Conclusion

In this paper, we have proposed LARV that jointly reconfigures the functional splits of the BSs, the resources and placements of vDUs and vCUs, and the routing for each BS flow. The objective of LARV is to minimize the long-term total operation cost while adapting to the possibly-varying traffic demands and resource availability. In particular, we have analyzed the relations between the traffic demands and resource utilization in the vRAN system, which renders their relations have high variance and dependence on platform and platform load. We also have formulated a comprehensive cost model capturing the impacts of resource overprovisioning, instantiation and reconfiguration and the declined demands. We have developed LARV using a model-free deep RL paradigm to solve the sequential decision-making problem. The agent's neural network is developed using a combination of D3QN and action branching to tackle the large state space and multi-dimensional action space. We also have conducted a series of trace-driven evaluations during the training process and online operation. The numerical results have shown that LARV successfully learns the optimal policy, where its learning convergence can be expedited through transfer learning even in different vRAN systems. Moreover, LARV offers considerable cost savings by up to 59% of the static benchmark, 35% of DDPG with discretization, and 76% of a distributed non-branching D3QN solution.

The proposed framework in this paper has been evaluated in a realistic simulated vRAN system based on collected testbed traces and network datasets. However, it has not been implemented in a real live network due to the limitation of the current testbed setup, i.e., it could not support several functional splits and the geographical location of the servers. In the future, implementing the framework and evaluating its performance in a real live network setup would be an interesting study.

References

[1] L. Bonati, M. Polese, S. D'Oro, S. Basagni, and T. Melodia, "Open, programmable, and virtualized 5G networks: State-of-the-art and the road ahead," *Comput. Netw.*, vol. 182, Dec. 2020, Art. no. 107516.

[2] "5G immersive service opportunities with edge cloud and cloud RAN," Nokia, Espoo, Finland, White Paper, 2019.

[3] "Nokia mobile anyhaul," Nokia, Espoo, Finland, White Paper, 2017.

[4] "Open and virtualized—The future radio access network," NEC, Tokyo, Japan, 2020.

[5] *O-RAN-WG1-O-RAN Architecture Description V08.00*, O-RAN Alliance, Alfter, Germany, 2023.

[6] *Architecture Description (Release 16), Version 16.1.0*, 3GPP Standard (NG-RAN) 38.401, 2020.

[7] F. W. Murti, J. A. Ayala-Romero, A. Garcia-Saavedra, X. Costa-Pérez, and G. Iosifidis, "An optimal deployment framework for multi-cloud virtualized radio access networks," *IEEE Trans. Wireless Commun.*, vol. 20, no. 4, pp. 2251–2265, Apr. 2021.

[8] G. Paschos, E. Bastug, I. Land, G. Caire, and M. Debbah, "Wireless caching: Technical misconceptions and business barriers," *IEEE Commun. Mag.*, vol. 54, no. 8, pp. 16–22, Aug. 2016.

[9] A. M. Alba, J. H. G. Velásquez, and W. Kellerer, "An adaptive functional split in 5G networks," in *Proc. IEEE Conf. Comput. Commun. Workshops (INFOCOM WKSHPS)*, 2019, pp. 410–416.

[10] J. A. Ayala-Romero, A. Garcia-Saavedra, M. Gramaglia, X. Costa-Pérez, A. Banchs, and J. J. Alcaraz, "vrAIn: Deep learning based orchestration for computing and radio resources in vRANs," *IEEE Trans. Mobile Comput.*, vol. 21, no. 7, pp. 2652–2670, Jul. 2022.

[11] J. A. Ayala-Romero, A. Garcia-Saavedra, X. Costa-Perez, and G. Iosifidis, "Orchestrating energy-efficient vRANs: Bayesian learning and experimental results," *IEEE Trans. Mobile Comput.*, vol. 22, no. 5, pp. 2910–2924, May 2023.

[12] I. Gomez-Miguelez, A. Garcia-Saavedra, P. D. Sutton, P. Serrano, C. Cano, and D. J. Leith, "SrsLTE: An open-source platform for LTE evolution and experimentation," in *Proc. 10th ACM Int. Workshop Wireless Netw. Testbeds, Exp. Eval. Characterization*, 2016, pp. 25–32.

[13] J. A. Ayala-Romero, A. Garcia-Saavedra, X. Costa-Perez, and G. Iosifidis, "EdgeBOL: Automating energy-savings for mobile edge AI," in *Proc. 17th Int. Conf. Emerg. Netw. Exp. Technol.*, 2021, pp. 397–410. [Online]. Available: https://doi.org/10.1145/3485983.3494849

[14] X. Foukas and B. Radunovic, "Concordia: Teaching the 5G vRAN to share compute," in *Proc. ACM SIGCOMM Conf.*, 2021, pp. 580–596. [Online]. Available: https://doi.org/10.1145/3452296.3472894

[15] Z. Wang, T. Schaul, M. Hessel, H. Van Hasselt, M. Lanctot, and N. De Freitas, "Dueling network architectures for deep reinforcement learning," in *Proc. ICML*, 2016, pp. 1–9.

[16] H. V. Hasselt, A. Guez, and D. Silver, "Deep reinforcement learning with double Q-learning," in *Proc. AAAI*, 2016, pp. 2094–2100.

[17] A. Tavakoli, F. Pardo, and P. Kormushev, "Action branching architectures for deep reinforcement learning," in *Proc. AAAI Conf. Artif. Intell.*, 2018, pp. 4131–4138.

[18] A. Garcia-Saavedra, G. Iosifidis, X. Costa-Perez, and D. J. Leith,, "Joint optimization of edge computing architectures and radio access networks," *IEEE J. Sel. Areas Commun.*, vol. 36, no. 11, pp. 2433–2443, Nov. 2018.

[19] B. Ojaghi, F. Adelantado, and C. Verikoukis, "SO-RAN: Dynamic RAN slicing via joint functional splitting and MEC placement," *IEEE Trans. Veh. Technol.*, vol. 72, no. 2, pp. 1925–1939, Feb. 2023.

[20] F. Z. Morais et al., "PlaceRAN: Optimal placement of virtualized network functions in beyond 5G radio access networks," *IEEE Trans. Mobile Comput.*, early access, Apr. 29, 2022, doi: 10.1109/TMC.2022.3171525.

[21] A. M. Alba, S. Janardhanan, and W. Kellerer, "Enabling dynamically centralized RAN architectures in 5G and beyond," *IEEE Trans. Netw. Service Manag.*, vol. 18, no. 3, pp. 3509–3526, Sep. 2021.

[22] A. M. Alba and W. Kellerer, "Dynamic functional split adaptation in next-generation radio access networks," *IEEE Trans. Netw. Service Manag.*, vol. 19, no. 3, pp. 3239–3263, Sep. 2022.

[23] D. Harutyunyan and R. Riggio, "Flex5G: Flexible functional split in 5G networks," *IEEE Trans. Netw. Service Manag.*, vol. 15, no. 3, pp. 961–975, Sep. 2018.

[24] D. Bega, A. Banchs, M. Gramaglia, X. Costa-Pérez, and P. Rost, "CARES: Computation-aware scheduling in virtualized radio access networks," *IEEE Trans. Wireless Commun.*, vol. 17, no. 12, pp. 7993–8006, Dec. 2018.

[25] S. Ali et al., "6G white paper on machine learning in wireless communication networks," 2020, *arXiv:2004.13875*.

[26] S. Matoussi, I. Fajjari, N. Aitsaadi, and R. Langar, "Deep learning based user slice allocation in 5G radio access networks," in *Proc. IEEE 45th Conf. Local Comput. Netw. (LCN)*, 2020, pp. 286–296.

[27] M. Kalntis and G. Iosifidis, "Energy-aware scheduling of virtualized base stations in O-RAN with online learning," in *Proc. IEEE Global Commun. Conf.*, 2022, pp. 1–7.

[28] T. Pamuklu, M. Erol-Kantarci, and C. Ersoy, "Reinforcement learning based dynamic function splitting in disaggregated green open RANs," in *Proc. EEE Int. Conf. Commun.*, 2021, pp. 1–6.

[29] L. Bonati, S. D'Oro, M. Polese, S. Basagni, and T. Melodia, "Intelligence and learning in O-RAN for data-driven NextG cellular networks," *IEEE Commun. Mag.*, vol. 59, no. 10, pp. 21–27, Oct. 2021.

[30] M. Polese, L. Bonati, S. D'Oro, S. Basagni, and T. Melodia, "ColO-RAN: Developing machine learning-based xApps for open RAN closed-loop control on programmable experimental platforms," *IEEE Trans. Mobile Comput.*, early access, Jul. 4, 2022, doi: 10.1109/TMC.2022.3188013.

[31] F. W. Murti, S. Ali, and M. Latva-aho, "Constrained deep reinforcement based functional split optimization in virtualized RANs," *IEEE Trans. Wireless Commun.*, vol. 21, no. 11, pp. 9850–9864, Nov. 2022.

[32] F. W. Murti, S. Ali, G. Iosifidis, and M. Latva-Aho, "Learning-based orchestration for dynamic functional split and resource allocation in vRANs," in *Proc. Joint Eur. Conf. Netw. Commun. 6G Summit (EuCNC/6G Summit)*, 2022, pp. 243–248.
[33] M. Polese, L. Bonati, S. D'Oro, S. Basagni, and T. Melodia, "Understanding O-RAN: Architecture, interfaces, algorithms, security, and research challenges," *IEEE Commun. Surveys Tuts.*, vol. 25, no. 2, pp. 1376–1411, 2nd Quart., 2023.
[34] D. Bega, M. Gramaglia, M. Fiore, A. Banchs, and X. Costa-Perez, "AZTEC: Anticipatory capacity allocation for zero-touch network slicing," in *Proc. IEEE Conf. Comput. Commun.*, 2020, pp. 794–803.
[35] *5G-CORAL: Refined Design of 5G-CORAL Orchestration and Control System and Future Directions, Public Deliverable D3.2*, Eur. Commission, Brussels, Belgium, May 2019.
[36] V. Mnih et al., "Human-level control through deep reinforcement learning," *Nature*, vol. 518, pp. 529–533, Feb. 2015.
[37] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. 3rd Int. Conf. Learn. Represent.*, 2015, pp. 1–15.
[38] T. Italia, "Telecommunications—SMS, Call, Internet—MI." 2015. [Online]. Available: https://doi.org/10.7910/DVN/EGZHFV
[39] B. Xiang, J. Elias, F. Martignon, and E. Di Nitto, "A dataset for mobile edge computing network topologies," *Data Brief*, vol. 39, Dec. 2021, Art. no. 107557.
[40] B. M. Waxman, "Routing of multipoint connections," *IEEE J. Sel. Areas Commun.*, vol. 6, no. 9, pp. 1617–1622, Dec. 1988.
[41] P. Rost, S. Talarico, and M. C. Valenti, "The complexity–rate tradeoff of centralized radio access networks," *IEEE Trans. Wireless Commun.*, vol. 14, no. 11, pp. 6164–6176, Nov. 2015.
[42] T. P. Lillicrap et al., "Continuous control with deep reinforcement learning," in *Proc. 4th Int. Conf. Learn. Represent.*, 2016, pp. 1–14.

Fahri Wisnu Murti received the B.S. degree from Telkom University, Indonesia, and the master's degree from the Department of IT Convergence Engineering, Kumoh National Institute of Technology, South Korea. He is currently pursuing the Ph.D. degree with the Centre for Wireless Communications, University of Oulu, Finland. Prior to his doctoral study, he has worked as a Research Assistant with the Department Computer Science, Trinity College Dublin, Ireland. His current research interests lie in the development of machine learning and optimization techniques for intelligent wireless networks.

Samad Ali received the B.S. degree in electrical engineering from the University of Tabriz, Tabriz, Iran, and the M.S. and Ph.D. degrees in wireless communications engineering from the University of Oulu, Oulu, Finland. He is currently a Postdoctoral Researcher with the University of Oulu and a Senior Research Specialist with Nokia Standards. His research interest is the applications of AI/ML in wireless communication systems.

George Iosifidis received the Diploma degree in electronics and telecommunications engineering from Greek Air Force Academy, Athens, in 2000, and the Ph.D. degree from the University of Thessaly in 2012. He was an Assistant Professor with Trinity College Dublin from 2016 to 2020 and a Research Scientist with Yale University from 2015 to 2017. He is currently an Associate Professor with the Delft University of Technology. His research interests lie in the broad area of network optimization and economics; more information can be found at www.FutureNetworksLab.net.

Matti Latva-aho (Senior Member, IEEE) received the M.Sc., Lic.Tech., and Dr.Tech. (Hons.) degrees in electrical engineering from the University of Oulu, Finland, in 1992, 1996, and 1998, respectively. From 1992 to 1993, he was a Research Engineer with Nokia Mobile Phones, Oulu, Finland, after which he joined the Centre for Wireless Communications (CWC), University of Oulu. He was the Director of CWC from 1998 to 2006 and the Head of the Department for Communication Engineering until August 2014. He is currently a Professor of Wireless Communications with the University of Oulu and the Director for National 6G Flagship Programme. He is also a Global Fellow with Tokyo University. He has published over 600 conference or journal papers in the field of wireless communications. He received the Nokia Foundation Award in 2015 for his achievements in mobile communications research.