IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 21 July 2024, accepted 1 August 2024, date of publication 6 August 2024, date of current version 20 August 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3439257

RESEARCH ARTICLE

Enhancing Cost-Effective 5G Virtualized RAN Pooling Gain on Clouds: An Intelligent Auto-Scaling Approach

KIHO CHO[1], JOONGHEON KIM[2], (Senior Member, IEEE), AND SOYI JUNG[3], (Member, IEEE)

[1]Samsung Electronics Network Business, Samsung Electronics, Suwon 16677, Republic of Korea
[2]Department of Electrical of Computer Engineering, Korea University, Seoul 02841, Republic of Korea
[3]Department of Electrical of Computer Engineering, Ajou University, Suwon 16499, Republic of Korea

Corresponding authors: Joongheon Kim (joongheon@korea.ac.kr) and Soyi Jung (sjung@ajou.ac.kr)

This work was supported by the National Research Foundation of Korea (NRF) Grant funded by the Korea Government [Ministry of Science and ICT (Information and Communications Technology) (MSIT)] under Grant RS-2024-00358662.

ABSTRACT Cloud computing has revolutionized information technology (IT) infrastructure by providing benefits such as scalability, flexibility, and convenience. If the radio access networks of wireless systems are constructed on edge clouds, operators can obtain various advantages. One notable benefit is pooling gain achieved through auto-scaling, leading to savings on investment. However, conventional auto-scaling methods employ numerous instances to prevent traffic overloads and ensure reliable system performance during peak demand. This traditional approach increases the burden on computing resources. To address this, this paper introduces an intelligent auto-scaling scheme that enhances pooling gain by leveraging the distinctive attributes of cloud computing. Additionally, we refine pooling gain to measure the performance of the fifth generation (5G) base stations on edge clouds. The simulation outcomes of our advanced auto-scaling scheme demonstrate its ability to achieve substantial improvements in productivity compared to the simpler implementations used in distributed cell sites, both in terms of pooling gain and the efficient utilization of computing resources.

INDEX TERMS Cloud computing, 5G virtualized base station, pooling gain, intelligent auto-scaling.

I. INTRODUCTION

A. EMERGING CLOUD COMPUTING

Cloud computing business has been growing fast last decade as it was able to innovate information technology (IT) infrastructure using cloud data centers [1], [2]. Especially large-scale public cloud companies called hyperscalers have been successful. They offer massive computing resources and especially enable hyperscale computing. Major hyperscalers are Amazon Web Service (AWS), Google Cloud Platform (GCP), Microsoft Azure [3], [4]. Users use large-scale cloud computing like a hyperscaler to deploy and operate large-scale applications and have significant benefits like scalability, flexibility, reliability, and convenience compared to on-premises computing systems [5], [6], [7]. Users can flexibly pull required computing resources according to varying demands [8]. Abundant computing resource of hyperscalers provides natural fault tolerance [9], [10], [11]. Hyperscalers can also provide a convenient working environment in that users do not have to manage on-premises computing systems by themselves, and they continue to provide up-to-date tools and application programming interface (API) of various purposes [12]. These are reasons why global IT infrastructure has been transformed and moved to hyperscalers.

B. HYBRID CLOUD COMPUTING FOR RADIO ACCESS NETWORK

In the telecommunication industry, leading operators are already, at least partly, using hyperscalers [13], [14].

The associate editor coordinating the review of this manuscript and approving it for publication was Ding Xu.

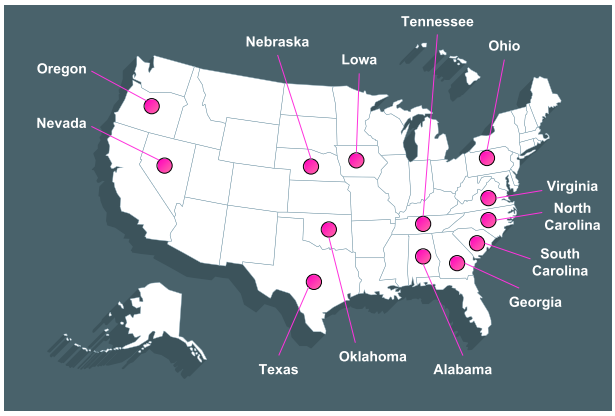


FIGURE 1. Location of data centers of Google Cloud Platform in USA [16].

Hyperscalers operate geographically distributed large-scale data centers interconnected via a high-speed backbone network as shown in Fig. 1 [15], [16]. Therefore, core network functions and management functions of wireless network operators can be easily relocated to nearby hyperscaler's data centers if operators take a risk to move their key information assets outside the company.

On the radio access network (RAN) side, it is possible to implement RAN by using hybrid cloud computing, which utilizes edge clouds as well as hyperscalers together when RAN is virtualized [17]. Edge clouds are cloud computing operated at the edge of the network, that is, close to base stations or user devices. They are considered necessary, especially for the fifth generation (5G) networks of which the major application is machine-to-machine communication to require low latency responses. As illustrated in Fig. 1, data centers of hyperscalers are centralized and not uniformly distributed. On the other hand, RAN is a network of many base stations of which each takes charge of radio transmissions in a small cell coverage called a cell. Base stations are uniformly distributed across the nation to provide nationwide wireless service. Thus, hybrid cloud computing utilizing edge clouds is necessary to implement virtualized RAN. When edge clouds are used for RAN, it should be considered that virtualized RAN software has real-time requirements according to 5G standards [18], [19]. All the base stations in the cellular network are synchronized and transmit and receive in a slotted pattern. Thus, physical and media access control (MAC) layers of base stations have to finish their jobs before the next slots start. Because of this real-time characteristic, each task in RAN software needs the amount of computing resources enough to process the peak rate of bursty data traffic. Otherwise, delayed traffic blocks would be lost and not transmitted. As described in [20], the pooling gain of virtualized RAN is one of the main factors in determining the gain of RAN architecture and the effect of edge clouds. In a centralized RAN architecture using edge clouds, commercial off-the-shelf (COTS) servers required to implement virtualized base stations are located in edge clouds. In distributed RAN architecture, servers are distributed to cell sites. We can expect that significant pooling gain can be achieved in a centralized RAN using edge clouds because multiple servers are shared by real-time tasks of 5G base stations. As well known in queueing theory [21], a fast single server can have better performance than multiple slow servers.

In this paper, we study the implementation of 5G virtualized base stations on servers and propose four different ways of implementation, including auto-scaling schemes on edge clouds. Furthermore, we redefine pooling gain based on the productivity concept and analyze how much pooling gain can be achieved for each implementation using a simple simulation model of 5G base stations. Through such analysis, the experimental results substantiate that our proposed intelligent auto-scaling technique can yield advantages in terms of computing cost compared to conventional methods.

C. CONTRIBUTIONS

The contributions of this paper are summarized as follows.

- This paper proposes various implementation approaches and auto-scaling mechanisms that are suitable for edge clouds (i.e., 5G virtualized base stations).
- This research demonstrates that the proposed auto-scaling scheme utilizing the intelligent load balancer yields a significant pooling gain compared to the conventional simple pooling scheme of 5G RAN deployed in distributed cell sites.
- The proposed strategy of this paper has the capability to automatically optimize resource utilization across varying load conditions, demonstrating a flexible and adaptive response to demand.

D. ORGANIZATION

The rest of this paper is organized as follows. Sec. II presents related works, and Sec. III furnishes technical insights into the merits, architecture, and operational mechanisms of the edge cloud within the context of 5G RAN. Next, Sec. IV redefines the pooling gain and formulates a structural model for 5G virtualized RAN implementation. In addition, Sec. V evaluates the performance of the proposed 5G virtualized RAN model. Lastly, Sec. VI provides various discussions, and Sec. VII concludes this paper.

II. RELATED WORK

The integration of 5G networks with edge cloud technologies has the rapid evolution and deployment of IT [22], [23], [24], [25]. With this trend, a wealth of research and development in the telecommunications sector are spurred, including the architecture, benefits, and challenges of 5G virtualized base stations [26], [27], [28], [29]. These studies have underscored the importance of efficient resource allocation, especially when aiming to achieve optimal peak rates in dynamic network environments [30], [31]. In this context, integrating edge clouds for 5G RAN can be a promising solution, which has the potential advantages for cost savings

based on pooling gains [32], [33], [34], [35]. Auto-scaling, a topic that has garnered significant attention, is another crucial component in the 5G landscape [36], [37]. However, traditional auto-scaling methods indiscriminately distributes input traffic. Furthermore, the role of load balancers in 5G networks is also imperative [38], [39]. As networks grow more complex and the demand for seamless connectivity increases, the need for enhanced load balancers becomes paramount [32], [40]. These advanced systems, capable of identifying specific processing units and reallocating them as needed, are required to ensure balanced instance loads and efficient network operations. Lastly, the network and source slicing related research results were also essential for our 5G virtualizad RAN network optimization and efficient management [41], [42].

In a nutshell, while the benefits of integrating 5G virtualized base stations with edge clouds are numerous, the challenges are equally daunting. Therefore, this paper offers novel insights and methodologies that can optimize 5G RAN implementations on edge clouds.

III. TECHNICAL BACKGROUNDS

A. BENEFITS OF USING EDGE CLOUDS

Wireless network operators can benefit from using hybrid cloud computing or edge clouds for their 5G RAN as follows:

- **Offloading the burden of managing complicated infrastructure.** Currently, operators own their infrastructure and need to manage all the equipment used in their networks to provide non-breaking communication service. Using edge clouds, that is, as many parts of RAN as possible are operating on edge clouds, operators can be relieved from managing infrastructure and more focused on 5G RAN application software because edge clouds are managed by companies like hyperscalers.
- **Saving investment with flexible allocation of computing resources.** Traditionally, a large-scale wireless network is built with a huge amount of special-purpose hardware to cover nationwide coverage and accommodate the highest demand over the year. Whenever operators upgrade networks with new technology, they have to replace most legacy products, which leads to a huge amount of investment. On the other hand, when 5G virtualized RANs are used and working on edge clouds, great investment saving is possible. Virtualized RAN software can operate on COTS servers, and software upgrade for new technology is possible without hardware change. Virtualization technology and plentiful computing resources in edge clouds allow operators to flexibly control the amount of used resources according to the varying demand. Pooling can also be applied to reduce the amount of required resources. Cloud companies generally adopt the policy of 'pay as you go [43].'
- **Reducing development and managing efforts with many application programming interfaces (APIs) and tools.** There are already many useful tools and APIs available in the virtualization and IT industry. When network equipment is based on special-purpose hardware, the diversity of useful tools verified in other IT areas must be limited. Virtualized RAN can improve chances to utilize them while developing and managing software.
- **Improving reliability.** When a fault happens in the network, virtualization technology in edge clouds allows faulty functions to be replaced with normal functions instantly, which results in improved reliability of the whole network.

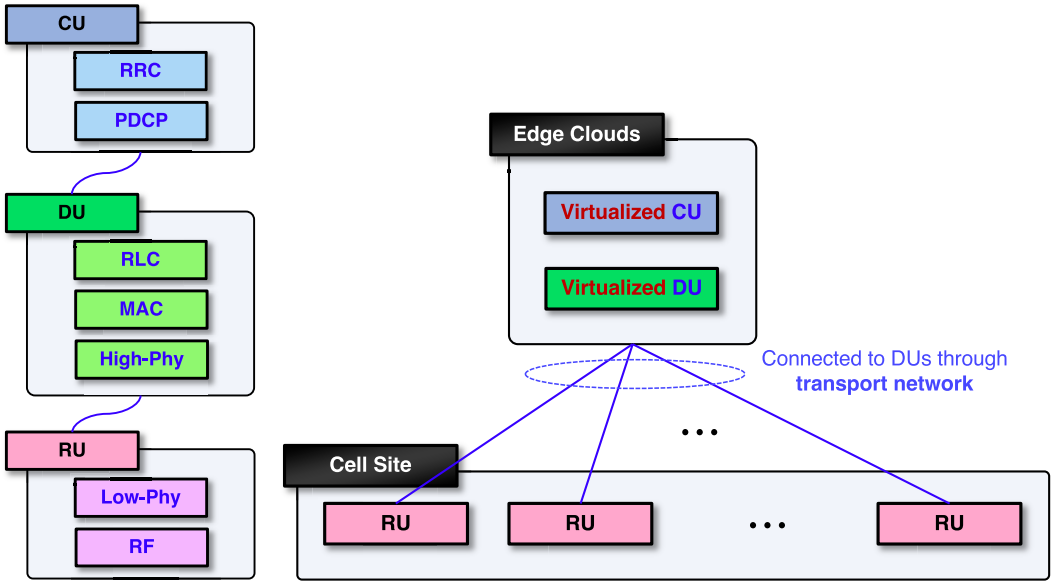
From the perspective of 5G operators, there are additional benefits, as follows:

- **Sharing network with other operators.** RAN sharing is one of the important topics for operators. It is sometimes a waste of investment that each operator has to equip their own network. It must be a better strategy for operators to share networks in rural areas with other operators and should focus on core business areas. When clouds are used and shared among multiple operators, RAN sharing must be much easier.
- **Well-structured software-defined network.** Current physical implementation of the 5G network is closely coupled with high-layer functional configuration. This implementation might be optimal in terms of product cost and network performance. But one of the disadvantages is that the network could need to be physically reconfigured even when there are some tweaks in high-layer functions. The resulting maintenance cost could increase much higher than initially expected. 5G virtualized RAN is a software-based implementation and independent from the below infrastructure. It can be easily configured and managed through software reconfiguration when it is constructed on well-structured edge clouds. Another disadvantage of traditional 5G RAN is to be difficult to equip and test the same configuration as commercial networks. There exist various configurations of special-purpose hardware according to commercial networks. Testing all of them is a consuming process. Separating 5G software from infra network surely relieves these testing burdens and can result in a more robust network.

Many benefits are expected by using edge clouds for 5G RAN as described above.

B. IDEAL 5G RAN ARCHITECTURE USING EDGE CLOUDS

5G standard aims for disaggregated RAN architecture which can be engineered to fit various environments and allow the construction of a network with multi-vendor units. Fig. 2 shows the functional splits of 5G base stations defined in the third generation partnership project (3GPP) [44], [45], [46] and its ideal architecture using edge clouds. 5G base station is generally composed of a central unit (CU), distributed unit (DU), and radio unit (RU). Because the effect of the virtualization can be maximized when as many parts as



(a) Functional split. (b) Ideal virtualized RAN architecture.

FIGURE 2. Functional splits of 5G base station and ideal RAN architecture.

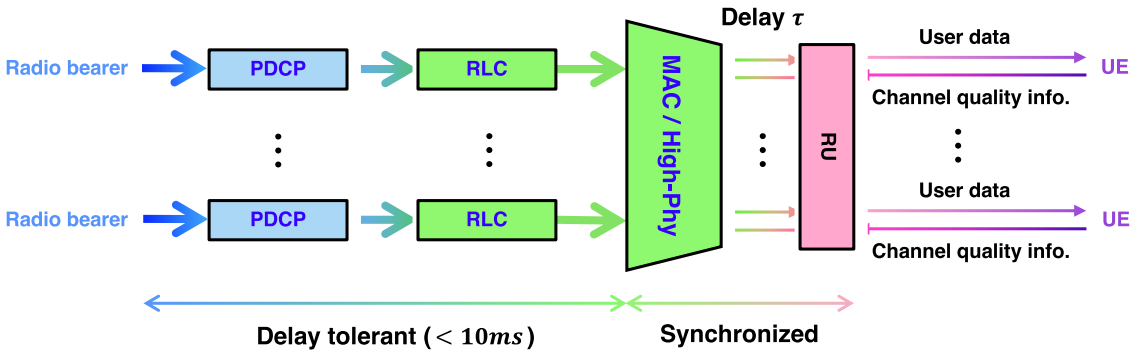


FIGURE 3. Operation of 5G RAN protocols.

possible are moved to edge clouds, CU and DU are located in edge clouds like ideal RAN architecture in Fig. 2 except RU, which is mainly composed of RF hardware. As shown in Fig. 2, CU and DU carry out radio resource control (RRC), packet data convergence protocol (PDCP), radio link control (RLC), MAC, and high-physical (High-Phy) layers of 5G RAN protocols. RUs are located in cell sites and connected to DUs through a transport network.

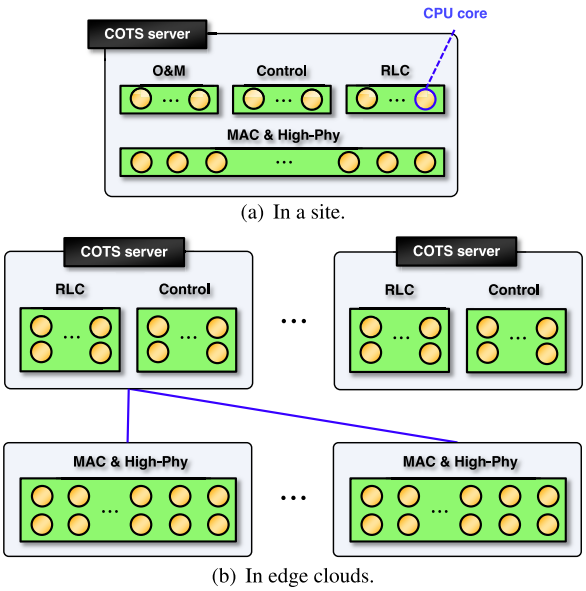
C. OPERATION OF 5G BASE STATIONS

The basic operation of 5G RAN protocols needs to be explained before stating the pooling gain of 5G RAN software. As shown in Fig. 3, upper layers like PDCP and RLC are composed of multiple entities for each radio bearer, and they are relatively delay tolerant. On the other hand, MAC and High-Phy layers are real-time operations [18], [19], [47]. MAC schedules which traffic in a cell is transmitted in the next slot depending on channel quality responses from UEs and traffic backlog in the buffer and then allocates selected traffic onto transport blocks. High-Phy layers carry out high-layer functions of the physical layer like modulation, pre-coding, and resource element mapping for the traffic of the next slot. RU takes charge of radio transmission and reception of a cell and has a one-to-one mapping with MAC and High-Phy layers of the cell. RUs are also synchronized with MAC and High-Phy layers.

IV. MODELING OF 5G VIRTUALIZED BASE STATIONS

A. POOLING GAIN AND PRODUCTIVITY OF SYSTEMS

Loading of RLC, MAC, and High-Phy layers can vary greatly in proportion to the traffic amount in a cell. Thus, the implementation of DU should be able to allocate to them the required computing resource to achieve the peak rate whenever it is required. Fig. 4 shows a typical core allocation in a COTS server for cells in a site, where enough cores should be allocated to guarantee the peak rate of each cell. In this allocation, the utilization ratio of computing resources could be low on average if peak-rate transmission occasionally



(a) In a site.

(b) In edge clouds.

FIGURE 4. Examples of central processing unit (CPU) core allocation for 5G RAN.

happens. Fig. 4 also shows different deployments of virtualized DU on edge clouds. Multiple servers can be used for cells in a cluster of sites. Many cores are allocated to and shared by more RLC, MAC, and High-Phy layers, which can generally improve the utilization ratio of servers thanks to statistical multiplexing among cells.

This kind of resource sharing is called 'pooling'. When a pooling system has pooling of some resource and it has the same performance as the original system, pooling gain can be defined as follows:

$$\text{Pooling gain} = 1 - \frac{R_{\text{pooling system}}}{R_{\text{original system}}}, \tag{1}$$

where R means resource amount. Because this definition can work only when both systems have the same performance, we extend the definition using the productivity concept. The productivity P of a system is defined as follows:

$$P = \frac{\text{Output}}{\text{Resource amount}} = \frac{O}{R}, \tag{2}$$

where O means the output of a system. Thus, pooling gain can be redefined as follows:

$$\text{Pooling gain} = 1 - \frac{P_{\text{original system}}}{P_{\text{pooling system}}}. \tag{3}$$

It can be observed that when the outputs of both systems are equal, this definition of pooling gain is equal to the previous definition.

B. MODELING OF MAC AND HIGH-PHY

A simple model to simulate the operation of 5G RAN can be made to evaluate the pooling gain and productivity of 5G RAN implementation as shown in Fig. 5. Each instance is a pod on edge clouds to run cell entities, including each cell's

TABLE 1. Parameter settings for simulation.

Notation		Value
N	Number of instances	30
n	Number of cell entities	6
r_{peak}	Resource for peak rate	100
r_{idle}	Resource when idle	10
r_{grt}	Minimum allocated resource	30
φ	A ratio of cell entity	0.5
k	Number of users in a cell	4
c_{avr}	Average cell capacity	700 [Mbps]
c_{grt}	Guaranteed cell capacity	200 [Mbps]
c_{peak}	Peak cell capacity	1 [Gbps]
σ	Standard deviation of cell capacity	100 [Mbps]
m	Burst size	2 [Mbyte]
B	Maximum buffer size	8 [Mbyte]

RAN protocol. Each instance has capacity according to the number of cores. There are N instances in the model. In an instance, n cell entities are allocated for pooling. All cell entities are synchronized with the primary clock. Cell entity e has an input buffer to save incoming traffic and process buffered traffic, and transmit to RU at the rate of output $b_{e,t}$, in slot t like the operation of 5G base stations.

In this model, cell capacity $c_{e,t}$ of cell entity e is the max capacity of the cell in slot t. Because the max capacity of a cell in each slot can vary according to the channel quality of UEs selected by a MAC scheduler, we simply generate it from $N(c_{\text{avr}}, \sigma^2)$ within the interval $(c_{\text{grt}}, c_{\text{peak}})$ in every slot. c_{avr}, c_{grt}, and c_{peak} mean average cell capacity, guaranteed cell capacity, and cell peak capacity, respectively. Guaranteed cell capacity is the cell capacity operators want to support at least in most cases when cells are deployed.

When cell entities process RAN protocols, they consume computing resources of instances. Part of the required computing resource is in proportion to the traffic rate. The other part is in proportion to the amount of air resources for wireless transmission. Because 5G uses adaptive modulation, though the traffic rate is not the peak rate, all the air resources can be used. We can assume that all the air resources are used to achieve cell capacity $c_{e,t}$ in slot t. Thus, for output rate $b_{e,t}$, the required computing resource ratio $v_{e,t}$ can be calculated as follows:

$$v_{e,t} = \frac{\varphi \cdot b_{e,t}}{c_{\text{peak}}} + \frac{(1-\varphi) \cdot b_{e,t}}{c_{e,t}}, \tag{4}$$

where φ is a ratio of the previous two parts. Cell entities need computing resources of r_{peak} to process traffic amount of c_{peak}; and they consume r_{idle} when no traffic. Therefore, for the required computing resource $r_{e,t}$ of cell entity e when the output rate is $b_{e,t}$ in slot t, we can use the following equation.

$$r_{e,t} = r_{\text{peak}} \cdot \left(v_{e,t} \cdot \left(1 - \frac{r_{\text{idle}}}{r_{\text{peak}}} \right) + \frac{r_{\text{idle}}}{r_{\text{peak}}} \right). \tag{5}$$

According to the required computing resources, instances allocate computing resources to each cell entity in a round-robin (RR) manner. When the resource is not enough, cell entities are allocated at least the minimum resource r_{grt} to

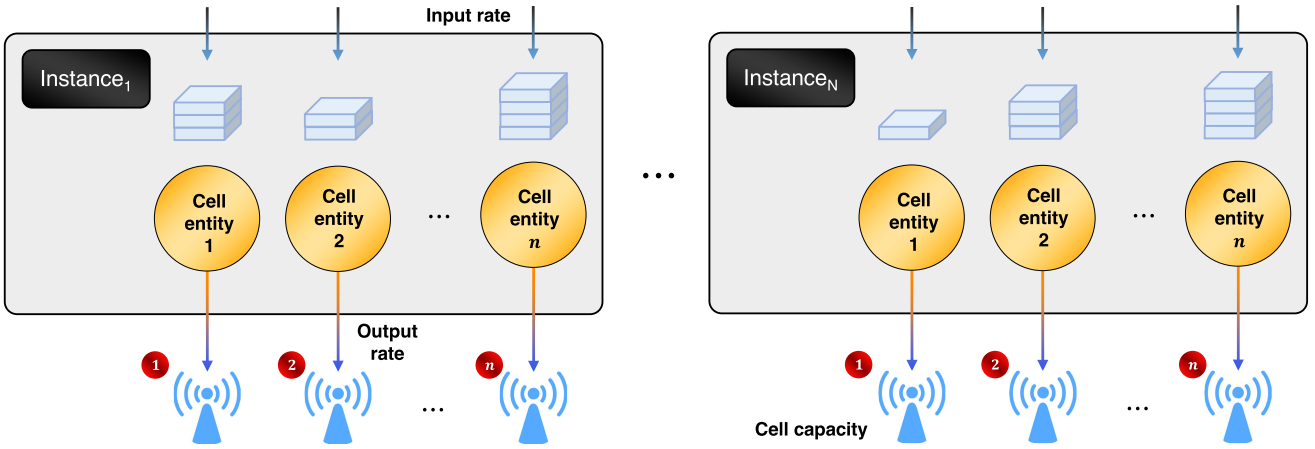


FIGURE 5. Modeling of 5G virtualized DU.

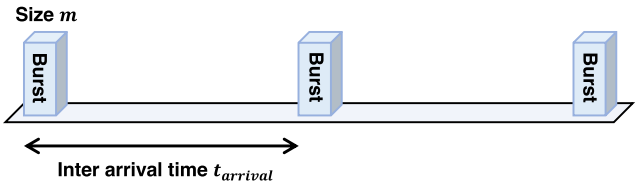


FIGURE 6. Input traffic model.

support the minimum rate. The total allocated computing resource in an instance is always less than instance capacity i, which is mathematically represented as $i \leq \sum_{e}^{n} r_e$.

Each cell entity has k users, which is the number of simultaneous users in each slot. Generally, this number is small when the traffic pattern is bursty. Each user generates input traffic which follows the on-off traffic model as shown in Fig. 6. Burst size is m, and inter-arrival time between bursts is a random variable $t_{\text{arrival}} \sim e^{c_{\text{avr}} \cdot \rho / k}$, where ρ is the loading of a cell. The input buffer size of each cell entity is kept less than the maximum buffer size B using flow control which backs off the next burst by another time t when the buffer is full.

For the simulation, parameters in Sec. IV-B can be set from the experience of a real 5G network as shown in Table 1.

C. ALGORITHM PSEUDO-CODES

In order to clearly present the differences between our proposed intelligent auto-scaling and conventional auto-scaling, the algorithm pseudo-codes are provided as following Algorithm 1 and Algorithm IV-C, respectively. As clearly stated in Algorithm 1, our proposed intelligent auto-scaling has the logic for identify whether the load of new instance is less than the average load.

Algorithm 1 Proposed Intelligent Auto-Scaling Scheme

```
Monitor average load among instances;
    IF overload in an instance happen;
        Create a new instance;
        FOR load of new instance < average load;
            DO relocate cell entities from loaded instance to the new instance;
            WHILE load of loaded instance < average load;
                Find next loaded instance > average load;
Return;
```

Algorithm 2 Conventional Auto-Scaling Scheme

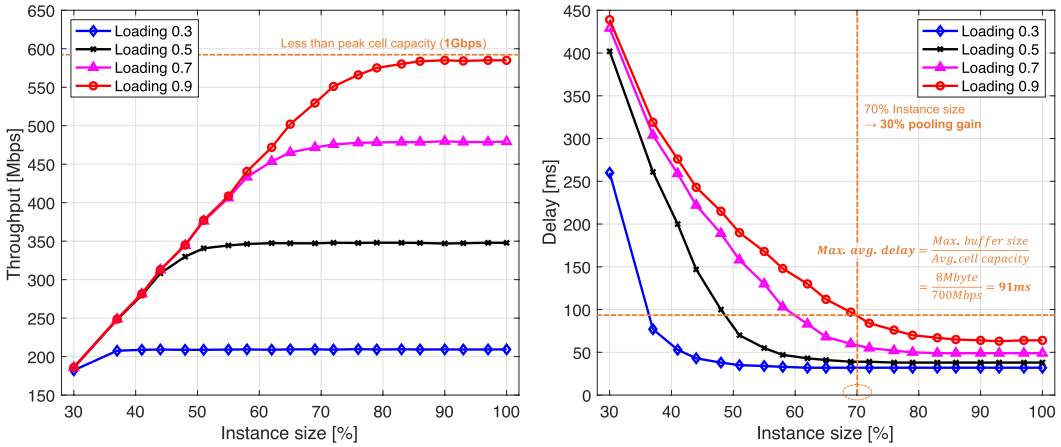
```
Monitor average load among instances;
    IF overload in an instance happen;
        Create a new instance;
        Relocate half of cell entities from the overload instance to the new instance;
Return;
```

V. PERFORMANCE RESULTS ACCORDING TO ALLOCATION SCHEMES

A. SIMPLE ALLOCATION SCHEME USED IN DISTRIBUTED CELL SITES

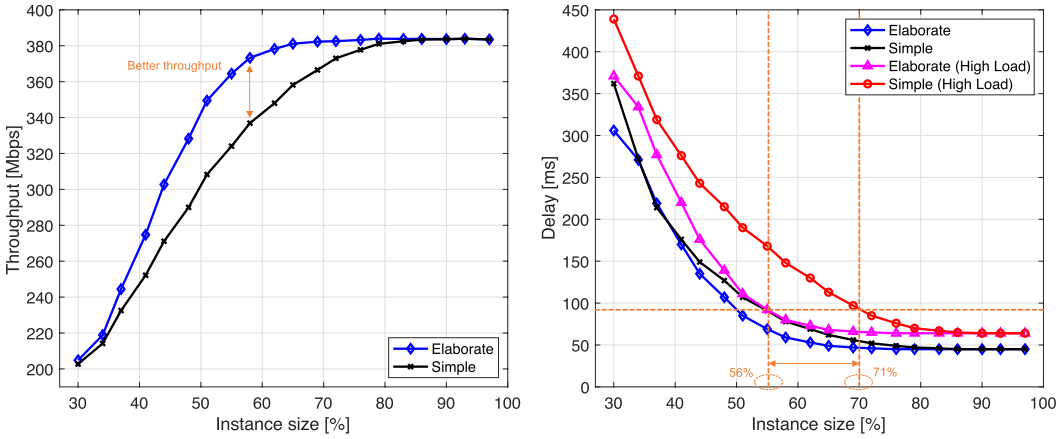
In this allocation, servers used for 5G base stations are assumed to be located in distributed cell sites, and their numbers are to be limited to reduce cost. Thus, cell processing entities for cells of a site should be processed with the limited computing resource of a server. We can safely allocate 6 cells in one instance in this simulation, considering the capacity of a server and the complexity of 5G base stations. Instance size is computing resources allocated to an instance, and it can vary according to the capacity of a server. In this simulation, we monitor the performance changing it from the guaranteed size to the peak size. The guaranteed size is assumed to be only 30 % of the peak size to support the guaranteed throughput of each cell. The peak size is to allocate 100 % resource required to support the peak throughput of each cell.

Fig. 7 shows the average throughput and the average delay in the buffer of cell entities under various instance sizes and loadings. We can choose the appropriate instance size for 5G base stations from the above graphs. We know that the maximum average delay is approximately $91\,ms \left(= \frac{B}{c_{\text{avr}}} = \frac{8\,Mbyte}{700\,Mbps}\right)$ because the maximum buffer size



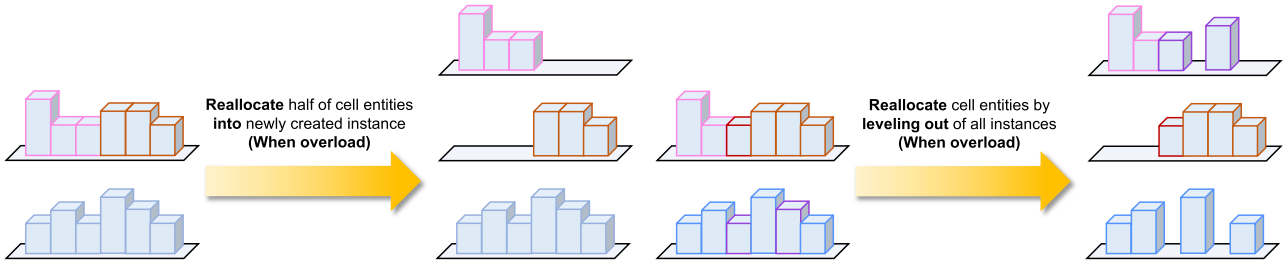
(a) Throughput vs. Instance size. (b) Delay vs. Instance size.

FIGURE 7. Average throughput and delay of cell entities under various instance sizes and loadings.



(a) Throughput in mixed loading. (b) Delay in mixed loading.

FIGURE 8. Throughput and delay of elaborate allocation in mixed loading.



(a) Simple load balancer. (b) Intelligent load balancer.

FIGURE 9. Two different auto-scaling schemes.

in this simulation is kept below 8Mbyte through flow control. Thus, 70 % instance size can be chosen from the delay curve of 90 % loading to meet this delay requirement. Thus, this implementation to use a simple pooling scheme can have 30 % pooling gain against the case simply using the peak size. Unfortunately, it is hard to say this size is optimal for every case. When most sites have less loading than 90 % for most of the operation time, cell entities do not need that big instance size. Even for 90 % loading, we can see that cell entities of this size cannot support the maximum throughput from the graph.

B. ALLOCATION METHODS FOR EDGE CLOUDS

1) ELABORATE ALLOCATION SCHEME

The cellular network generally has locality characteristics. During the daytime, business areas have high loading, and residential areas have low loading, but it is reversed during the night. Thus, if operators can mix cells of different localities

in an instance, they can have better pooling gain. In this simulation, we consider 3 types of cells with different input traffic loading. The first cell type is a dense urban cell with 90 % loading. 30 % of total cells are dense urban cells. Second one is urban cell with 60 % loading and they are 30 % of total cells. The third one is a rural cell with 30 % loading. Remain cells are all rural cells. Elaborate allocation is possible only in case base stations are big enough to cover different types of areas. Thus, this is possible when base stations are implemented as a centralized RAN architecture on edge clouds as described in Sec. IV-A.

Fig. 8 shows throughput and delay of elaborate allocation in the mixed loading situation. Elaborate allocation shows better throughput with smaller instance sizes. In terms of delay in buffer, it also has a shorter delay than simple allocation. Fig. 8(b) shows that the delay of high-load cells of Elaborate allocation is much shorter than that of Simple allocation. With elaborate allocation, we can choose 56 % instance size to meet the maximum delay of high load cells, which means elaborate allocation on edge cloud has 15 % more pooling gain against simple allocation in the mixed loading situation.

2) AUTO-SCALING: SIMPLE LOAD BALANCER

Though Elaborate allocation has pooling gain against Simple allocation, it is not always possible to find and mix cells of opposite loading characteristics. An advanced allocation scheme to adapt to various loading conditions automatically is required to achieve pooling gain in any circumstances. Hyperscalers provide auto-scaling feature for instances, which automatically increases the number of instances when overloading in an instance happens for a threshold time. We can use this feature to optimize instance size in any loading situation. First, we can come up with a simple load balancer to reallocate half of the cell entities in the overloaded instance into newly created instance when auto-scaling is applied, as shown in Fig. 9. In this simulation, auto-scaling initiates when a load of instance is over 70 % for 20 seconds continuously.

3) AUTO-SCALING: INTELLIGENT LOAD BALANCER

We can come up with an intelligent approach for auto-scaling. As shown in Fig. 9(b), an intelligent load balancer levels out the loading of all instances by reallocating cell entities when auto-scaling is applied. It has average loading information across all instances and reduces the loading of overloaded instances to an average level by reallocating cell entities in them to a new instance until its loading is close to average.

Auto-scaling schemes can increase the number of instances whenever overload happens. As a result, the maximum throughput and the minimum delay can be achieved even with a small instance size, as shown in Fig. 10. Instead, the number of instances can increase through scaling out, as shown in Fig. 11(a). We can see that delays of high load cells of auto-scaling schemes in Fig. 10(b) decrease as the number of instances increases with smaller instance size.

Fig. 11(a) shows that auto-scaling happens when instance size is less than 70 %. Smaller instance sizes need a greater number of instances. An intelligent load balancer generates a smaller number of instances than a simple load balancer. Fig. 11(b) shows productivity results calculated with throughput/total amount of instances as defined in Sec. III-A. Smaller instance sizes tend to bring in higher productivity. When it comes to the minimum size of the instance, the appropriate minimum size is to support one peak rate and $(n-1)$ guaranteed rates at the same time because it is the minimum size to support the peak rate of a cell in any cases. Since the number of cell entities is set to 6 in Table 1, the appropriate minimum size for our scenario is $\frac{100\,\%+30\,\%\times 5}{600\,\%} = \frac{250\,\%}{600\,\%} \simeq 0.42$ of the peak size which is the size to support all peak rates at the same time. Thus, 42 % instance size can be chosen for auto-scaling schemes.

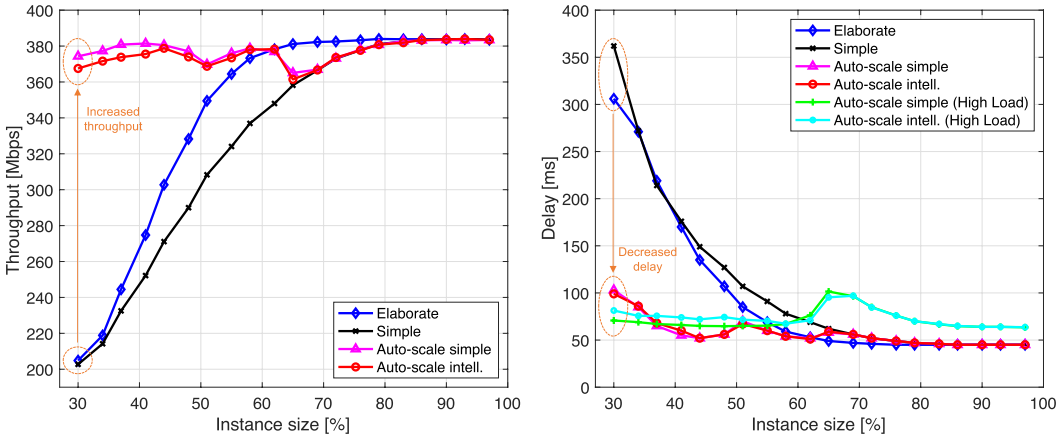
Fig. 12(a) shows productivities defined in Sec. IV-A according to loading. When loading is smaller than 40 %, auto-scaling schemes show much better productivity than simple allocation, which uses 70 % instance size. But a simple load balancer shows worse productivity after 60 % loading because the number of instances due to scaling out is too many. The productivity of an intelligent load balancer becomes worse than Simple allocation only when loading is 90 %. Fig. 12(b) shows the resulting pooling gain against the simple allocation. The simple load balancer cannot have positive pooling gain when loading is more than 50 %. Accordingly, the intelligent load balancer can have good pooling gain except in very high-loading cases.

If each loading is distributed equally, then the simple load balancer has 11 % pooling gain on average, and the intelligent load balancer has 24 % pooling gain. As a result, the proposed intelligent load balancer has 13.1 % more positive pooling gain against simple allocation as represented in Fig. 13. That is, using auto-scaling and intelligent load balancer 5G base stations can have better pooling gain than in an elaborate allocation scheme without efforts for mixing cells.

VI. DISCUSSIONS

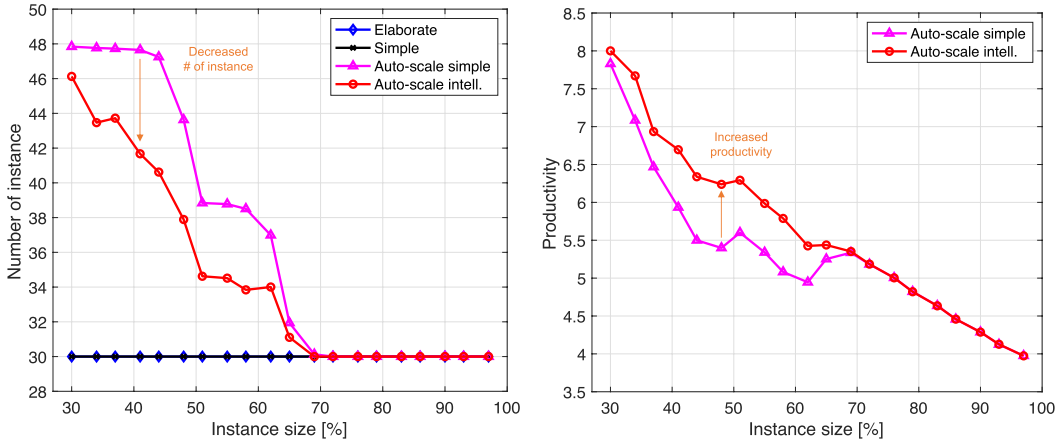
A. INSIGHT ON AUTO-SCALING FOR 5G BASE STATIONS

Through the above simulation study, we can acknowledge that it is important how to apply auto-scaling for the implementation of 5G base stations on edge clouds in terms of pooling gain. Current auto-scaling schemes that Hyperscalers generally provide are simple load balancers based on the assumption that input can be distributed to instances indiscriminately. So, load balancers just distribute input traffic equally to instances when an instance is newly created, as shown in Fig. 14. As we found in the case of 5G base stations, there can be applications in which inputs are destined for specific processing units. For these applications, the pooling of processing units in an instance is basically required, and auto-scaling scheme needs to be enhanced like an intelligent load balancer. It must be very useful that an



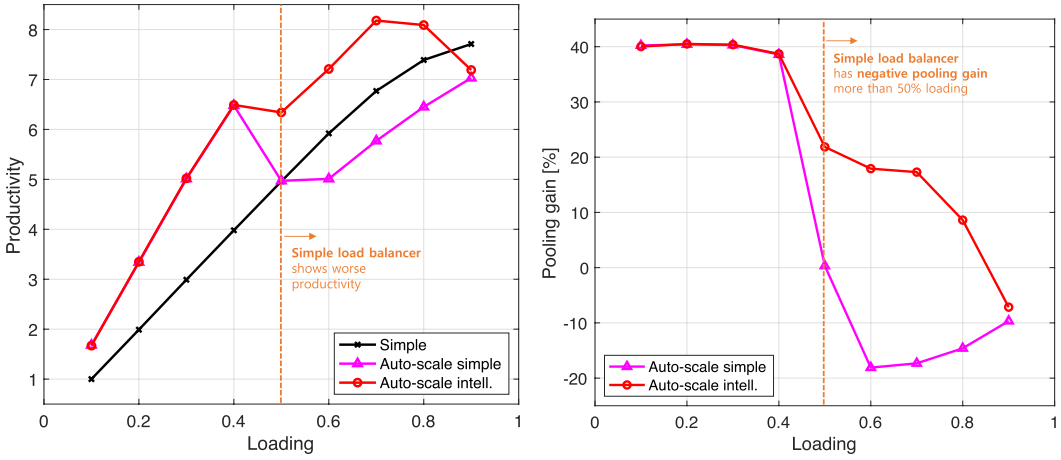
(a) Throughput of auto-scaling. (b) Delay of auto-scaling.

FIGURE 10. **Throughput and delay of auto-scaling schemes in a mixed loading situation.**



(a) Number of instances of auto-scaling in a mixed load. (b) Productivity of auto-scaling in mixed load.

FIGURE 11. **Number of instances and productivity of auto-scaling schemes.**



(a) Productivity. (b) Pooling gain against simple allocation.

FIGURE 12. **Productivity and pooling gain according to loading.**

enhanced load balancer is provided as one of auto-scaling schemes, which can identify processing units and reallocate them to instances to level out loadings of instances when an overload occurs.

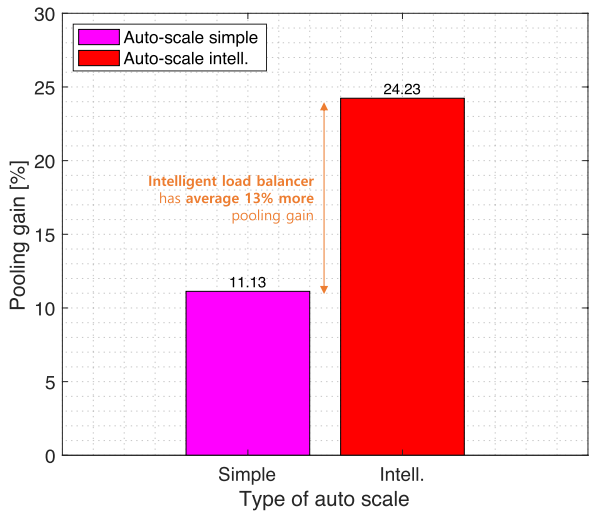
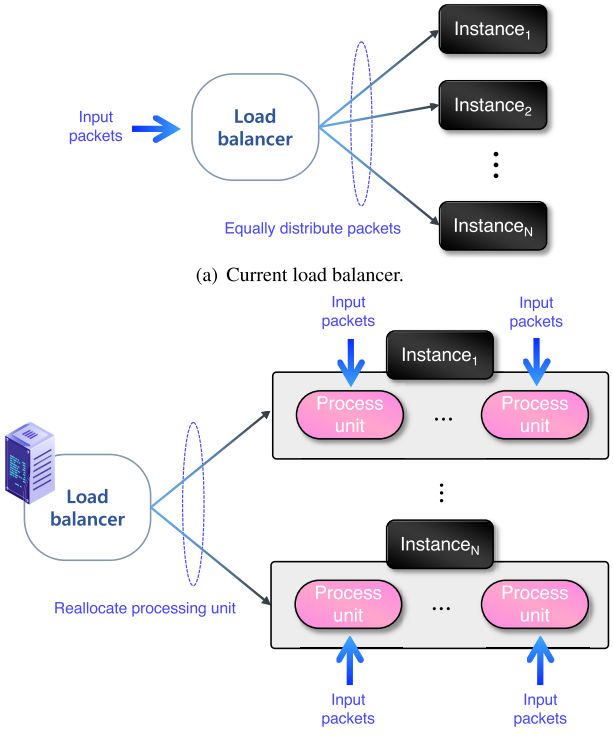


FIGURE 13. Average pooling gain in Fig. 12 according to auto-scaling scheme.



(a) Current load balancer.

(b) Enhanced load balancer.

FIGURE 14. Difference of current load balancer and enhanced load balancer to improve pooling gain.

B. BENEFITS OF EDGE CLOUDS FOR 5G RAN

It is necessary to highlight that the benefit of edge clouds is to enable the virtualized networks for 5G cellular networks. The edge clouds are located close to users and their benefit is to realize very fast responses to requests from users, which is a critical feature for 5G network because one of its promising features is machine-to-machine communications. The machine-type communications utilized for machine control require very fast responses beyond human responses. On the other hand, traditional centralized cloud usually is integrated in a few central sites, especially in large countries like USA and China as described in Sec. I and Fig. 1.

As described in Sec. III-B and Fig. 3, virtualized DUs in virtualized 5G networks are located in clouds. The DU and RU of 5G base-station has very strict delay requirements of 67 μs due to hybrid automatic repeat request (ARQ) operation, which has 3 ms time budget to the next transmission according to 5G standard and one of real implementations of base-stations can allow max 67 μs latency between them when excluding time for other operations like scheduling.

This latency of 67 μs allows about 20 km distance between RU and DU. The clouds of virtualized DUs also need to be not farther from RU sites than this distance and it is why edge clouds are needed for virtualized network. Generally, centralized clouds are way farther, and virtualized DU cannot be on centralized clouds mostly.

C. OPTIMAL RESOURCE USAGE

Each cell's traffic patterns and variations are various and intrinsic characteristics in each cell. It is almost impossible to forecast next input traffic exactly. Even in this case, the loss rate of data traffic must be kept very low. Thus, resource should be provided in a deterministic way. Especially, the guaranteed rate of each cell needs to be served at any time and the peak rate should to be served in all possible situations. Here, we can come up with the minimum resource allocation for each cell. Then, the instance size for multiple cell entities should be at least the amount enough to serve one peak rate and other guaranteed rates at any moment. In terms of server resources, we can state that this allocation is optimal if the withstanding low throughput of the network. Starting from this initial allocation, auto-scaling schemes monitor the loads of instances and find the bottleneck instance to limit cell throughput. When the overload in these instances happens, it adds new instance to resolve overload. Especially, because intelligent auto-scaling schemes can redistribute cell entities and level out, they can add minimal number of instances and we can state that it is close to optimal allocation as well as improved throughput.

VII. CONCLUSION AND FUTURE WORK

When edge clouds are utilized for 5G RAN, operators can have the benefits of virtualization that cloud computing delivers. Pooling gain is one of the important benefits in terms of cost saving. In this paper, we studied the architecture of 5G virtualized base stations and proposed various implementation approaches, including auto-scaling schemes available in edge clouds. Furthermore, we analyzed how much pooling gains can be achieved through simulation. Results of the simulation verify that our auto-scaling scheme employing an intelligent load balancer can bring 24 % pooling gain on average against a simple pooling scheme of 5G RAN used in distributed cell sites. Especially this scheme can automatically optimize resource usage in any loading situation. Thus, under low loading conditions, a maximum

40 % pooling gain is possible. Unlike current auto-scaling schemes, this study suggests the need for an enhanced auto-scaling scheme that can reallocate processing units across instances when scaling out happens.

As a potential future research direction, it is worthy to consider various and enhanced autho-scaling schemes in our considering cloud computing systems.

REFERENCES

[1] L. M. Dang, M. J. Piran, D. Han, K. Min, and H. Moon, "A survey on Internet of Things and cloud computing for healthcare," *Electronics*, vol. 8, no. 7, p. 768, Jul. 2019.

[2] H. M. Sabi, F.-M.-E. Uzoka, K. Langmia, and F. N. Njeh, "Conceptualizing a model for adoption of cloud computing in education," *Int. J. Inf. Manage.*, vol. 36, no. 2, pp. 183–191, Apr. 2016.

[3] P. Pierleoni, R. Concetti, A. Belli, and L. Palma, "Amazon, Google and Microsoft solutions for IoT: Architectures and a performance comparison," *IEEE Access*, vol. 8, pp. 5455–5470, 2020.

[4] A. Alsalemi, A. Al-Kababji, Y. Himeur, F. Bensaali, and A. Amira, "Cloud energy micro-moment data classification: A platform study," in *Proc. IEEE/ACM 13th Int. Conf. Utility Cloud Comput. (UCC)*, Leicester, U.K., Dec. 2020, pp. 420–425.

[5] S. Li, D. Niu, Y. Wang, W. Han, Z. Zhang, T. Guan, Y. Guan, H. Liu, L. Huang, Z. Du, F. Xue, Y. Fang, H. Zheng, and Y. Xie, "Hyperscale FPGA-as-a-service architecture for large-scale distributed graph neural network," in *Proc. 49th Annu. Int. Symp. Comput. Archit.*, New York, NY, USA, Jun. 2022, pp. 946–961.

[6] Capgemini. (Feb. 2021). *Hyperscalers and Managed Services: What's the Future*. [Online]. Available: https://www.capgemini.com/insights/expert-perspectives/hyperscalers-and-managed-services-whats-the-future/

[7] Y. Li, Z. M. Jiang, H. Li, A. E. Hassan, C. He, R. Huang, Z. Zeng, M. Wang, and P. Chen, "Predicting node failures in an ultra-large-scale cloud computing platform: An AIOps solution," *ACM Trans. Softw. Eng. Methodol.*, vol. 29, no. 2, pp. 1–24, Apr. 2020.

[8] T. Benjaponpitak, M. Karakate, and K. Sripanidkulchai, "Enabling live migration of containerized applications across clouds," in *Proc. IEEE Conf. Comput. Commun. (IEEE INFOCOM)*, Toronto, ON, Canada, Jul. 2020, pp. 2529–2538.

[9] T. Hoefler, A. Hendel, and D. Roweth, "The convergence of hyperscale data center and high-performance computing networks," *Computer*, vol. 55, no. 7, pp. 29–37, Jul. 2022.

[10] D. Saxena, I. Gupta, A. K. Singh, and C.-N. Lee, "A fault tolerant elastic resource management framework toward high availability of cloud services," *IEEE Trans. Netw. Service Manage.*, vol. 19, no. 3, pp. 3048–3061, Sep. 2022.

[11] L. Luo, S. Meng, X. Qiu, and Y. Dai, "Improving failure tolerance in large-scale cloud computing systems," *IEEE Trans. Rel.*, vol. 68, no. 2, pp. 620–632, Jun. 2019.

[12] R. Ramakrishnan et al., "Azure data lake store: A hyperscale distributed file service for big data analytics," in *Proc. ACM Int. Conf. Manage. Data (SIGMOD)*, Chicago, IL, USA, May 2017, pp. 51–63.

[13] A. Latif, A. Khamas, S. Goswami, V. P. Talari, and Y. Jung. (Feb. 2022). *Telco Meets AWS Cloud: Deploying Dish's 5G Network in AWS Cloud*. [Online]. Available: https://www.capgemini.com/insights/expert-perspectives/hyperscalers-and-managed-services-whats-the-future/

[14] AT&T. (Jul. 2021). *AT&T and Google Cloud Expand 5G and Edge Collaboration To Deliver Next-generation Bus. Outcomes*. [Online]. Available: https://about.att.com/story/2021/att_google_cloud.html

[15] B. Varghese, E. de Lara, A. Y. Ding, C.-H. Hong, F. Bonomi, S. Dustdar, P. Harvey, P. Hewkin, W. Shi, M. Thiele, and P. Willis, "Revisiting the arguments for edge computing research," *IEEE Internet Comput.*, vol. 25, no. 5, pp. 36–42, Sep. 2021.

[16] V. Sakalkar, "Hypersacle datacenter architecture trends," in *Proc. OCP Summit*, Nov. 2021, pp. 1–17.

[17] IBM. *What is Hybrid Cloud?* Accessed: Apr. 2021. [Online]. Available: https://www.ibm.com/topics/hybrid-cloud

[18] *NR; Physical Layer Procedures for Control*, document 3GPP TS 38.213, Version 16.3.0, Oct. 2020.

[19] *NR; Physical Layer Procedures for Data*, document 3GPP TS 38.214, Version 16.11.0, Sep. 2022.

[20] J. Liu, S. Zhou, J. Gong, Z. Niu, and S. Xu, "Statistical multiplexing gain analysis of heterogeneous virtual base station pools in cloud radio access networks," *IEEE Trans. Wireless Commun.*, vol. 15, no. 8, pp. 5681–5694, Aug. 2016.

[21] J. F. Shortle, J. M. Thompson, D. Gross, and C. M. Harris, *Fundamentals Queueing Theory*, vol. 399. Hoboken, NJ, USA: Wiley, 2018.

[22] K. Cao, Y. Liu, G. Meng, and Q. Sun, "An overview on edge computing research," *IEEE Access*, vol. 8, pp. 85714–85728, 2020.

[23] N. Hassan, K. A. Yau, and C. Wu, "Edge computing in 5G: A review," *IEEE Access*, vol. 7, pp. 127276–127289, 2019.

[24] X. Wang, G. Han, X. Du, and J. J. P. C. Rodrigues, "Mobile cloud computing in 5G: Emerging trends, issues, and challenges [guest editorial]," *IEEE Netw.*, vol. 29, no. 2, pp. 4–5, Mar. 2015.

[25] I. Ahmad, T. Kumar, M. Liyanage, J. Okwuibe, M. Ylianttila, and A. Gurtov, "Overview of 5G security challenges and solutions," *IEEE Commun. Standards Mag.*, vol. 2, no. 1, pp. 36–43, Mar. 2018.

[26] R. Vilalta, V. Lopez, A. Giorgetti, S. Peng, V. Orsini, L. Velasco, R. Serral-Gracia, D. Morris, S. De Fina, F. Cugini, P. Castoldi, A. Mayoral, R. Casellas, R. Martinez, C. Verikoukis, and R. Munoz, "TelcoFog: A unified flexible fog and cloud computing architecture for 5G networks," *IEEE Commun. Mag.*, vol. 55, no. 8, pp. 36–43, Aug. 2017.

[27] X.-Q. Pham, T.-D. Nguyen, T. Huynh-The, E.-N. Huh, and D.-S. Kim, "Distributed cloud computing: Architecture, enabling technologies, and open challenges," *IEEE Consum. Electron. Mag.*, vol. 12, no. 3, pp. 98–106, May 2023.

[28] M. Chen, Y. Zhang, Y. Li, S. Mao, and V. C. M. Leung, "EMC: Emotion-aware mobile cloud computing in 5G," *IEEE Netw.*, vol. 29, no. 2, pp. 32–38, Mar. 2015.

[29] R. Chaudhary, N. Kumar, and S. Zeadally, "Network service chaining in fog and cloud computing for the 5G environment: Data management and security challenges," *IEEE Commun. Mag.*, vol. 55, no. 11, pp. 114–122, Nov. 2017.

[30] B. Cao, Z. Sun, J. Zhang, and Y. Gu, "Resource allocation in 5G IoV architecture based on SDN and fog-cloud computing," *IEEE Trans. Intell. Transp. Syst.*, vol. 22, no. 6, pp. 3832–3840, Jun. 2021.

[31] L. Ferdouse, A. Anpalagan, and S. Erkucuk, "Joint communication and computing resource allocation in 5G cloud radio access networks," *IEEE Trans. Veh. Technol.*, vol. 68, no. 9, pp. 9122–9135, Sep. 2019.

[32] D. Wubben, P. Rost, J. S. Bartelt, M. Lalam, V. Savin, M. Gorgoglione, A. Dekorsy, and G. Fettweis, "Benefits and impact of cloud computing on 5G signal processing: Flexible centralization through cloud-RAN," *IEEE Signal Process. Mag.*, vol. 31, no. 6, pp. 35–44, Nov. 2014.

[33] M. E. Khoda, M. A. Razzaque, A. Almogren, M. M. Hassan, A. Alamri, and A. Alelaiwi, "Efficient computation offloading decision in mobile cloud computing over 5G network," *Mobile Netw. Appl.*, vol. 21, no. 5, pp. 777–792, Feb. 2016.

[34] T. X. Tran, A. Hajisami, and D. Pompili, "Cooperative hierarchical caching in 5G cloud radio access networks," *IEEE Netw.*, vol. 31, no. 4, pp. 35–41, Jul./Aug. 2017.

[35] M. Peng, Y. Li, Z. Zhao, and C. Wang, "System architecture and key technologies for 5G heterogeneous cloud radio access networks," *IEEE Netw.*, vol. 29, no. 2, pp. 6–14, Mar./Apr. 2015.

[36] C. Qu, R. N. Calheiros, and R. Buyya, "Auto-scaling web applications in clouds: A taxonomy and survey," *ACM Comput. Surv.*, vol. 51, no. 4, pp. 1–33, Jul. 2018.

[37] Y. Ren, T. Phung-Duc, J.-C. Chen, and Z.-W. Yu, "Dynamic auto scaling algorithm (DASA) for 5G mobile networks," in *Proc. IEEE Global Commun. Conf. (GLOBECOM)*, Washington, DC, USA, Dec. 2016, pp. 1–6.

[38] C. Tsai and M. Moh, "Load balancing in 5G cloud radio access networks supporting IoT communications for smart communities," in *Proc. IEEE Int. Symp. Signal Process. Inf. Technol. (ISSPIT)*, Bilbao, Spain, Dec. 2017, pp. 259–264.

[39] J. Jijin, B.-C. Seet, P. H. J. Chong, and H. Jarrah, "Service load balancing in fog-based 5G radio access networks," in *Proc. IEEE Int. Symp. Pers., Indoor, Mobile Radio Commun. (PIMRC)*, Montreal, QC, Canada, Oct. 2017, pp. 1–5.

[40] A. A. Abdelltif, E. Ahmed, A. T. Fong, A. Gani, and M. Imran, "SDN-based load balancing service for cloud servers," *IEEE Commun. Mag.*, vol. 56, no. 8, pp. 106–111, Aug. 2018.

[41] T. Kwantwi, G. Sun, N. A. E. Kuadey, G. Maale, and G. Liu, "Blockchain-based computing resource trading in autonomous multi-access edge network slicing: A dueling double deep Q-learning approach," *IEEE Trans. Netw. Service Manage.*, vol. 20, no. 3, pp. 2912–2928, Sep. 2023.
[42] G. Sun, G. O. Boateng, D. Ayepah-Mensah, G. Liu, and J. Wei, "Autonomous resource slicing for virtualized vehicular networks with D2D communications based on deep reinforcement learning," *IEEE Syst. J.*, vol. 14, no. 4, pp. 4694–4705, Dec. 2020.
[43] S. Zhao, J. Miao, J. Zhao, and N. Naghshbandi, "A comprehensive and systematic review of the banking systems based on pay-as-you-go payment fashion and cloud computing in the pandemic era," *Inf. Syst. e-Bus. Manage.*, vol. 22, pp. 1–29, Jan. 2023.
[44] *Study on New Radio Access Technology: Radio Access Architecture and Interfaces*, document 3GPP TR 38.801, Version 14.0.0, Mar. 2017.
[45] O. Andersson. (May 2021). *Functional Splits: The Foundation of an Open 5G RAN*. [Online]. Available: https://www.5gtechnologyworld.com/functional-splits-the-foundation-of-an-open-5g-ran/
[46] Ericsson, Huawei Technologies, NEC Corporation, and Nokia. (May 2019). *ECPRI Specification V2.0*. [Online]. Available: http://www.cpri.info/downloads/eCPRI_v_2.0_2019_05_10c.pdf
[47] *NR; Medium Access Control (MAC) Protocol Specification*, document 3GPP TS 38.321, Version 16.1.0, Jul. 2020.

KIHO CHO received the B.S., M.S., and Ph.D. degrees in computer science from Korea Advanced Institute of Science and Technology (KAIST), Daejeon, Republic of Korea, in 1993, 1995, and 1999, respectively.

He has been a corporate Vice President (VP) with the Network Business Unit, Samsung Electronics, Republic of Korea. He has participated in the design and development of 3G, 4G, and 5G mobile wireless networks, as a System Architect, after joining Samsung Electronics, in 1999. His current research interests include network virtualization on hyperscalers and network automation using artificial intelligence technology for 5G and beyond-5G networks.

JOONGHEON KIM (Senior Member, IEEE) received the B.S. and M.S. degrees in computer science and engineering from Korea University, Seoul, Republic of Korea, in 2004 and 2006, respectively, and the Ph.D. degree in computer science from the University of Southern California (USC), Los Angeles, CA, USA, in 2014.

He has been with Korea University, since 2019, where he is currently an Associate Professor with the Department of Electrical and Computer Engineering and also an Adjunct Professor with the Department of Communications Engineering (co-operated by Samsung Electronics) and the Department of Semiconductor Engineering (co-operated by SK Hynix). Before joining Korea University, he was a Research Engineer with LG Electronics, Seoul, from 2006 to 2009; a Systems Engineer with Intel Corporation Headquarter, Santa Clara, CA, from 2013 to 2016; and an Assistant Professor of computer science and engineering with Chung-Ang University, Seoul, from 2016 to 2019. He is the Executive Director of Korea Institute of Communication and Information Sciences (KICS). He was a recipient of the Annenberg Graduate Fellowship with his Ph.D. admission from USC, in 2009; the Intel Corporation Next Generation and Standards (NGS) Division Recognition Award, in 2015; the IEEE Systems Journal Best Paper Award, in 2020; the IEEE ComSoc Multimedia Communications Technical Committee (MMTC) Outstanding Young Researcher Award, in 2020; the IEEE ComSoc MMTC Best Journal Paper Award, in 2021; the Best Special Issue Guest Editor Award by *ICT Express*, in 2022; and the Best Editor Award by *ICT Express*, in 2023. He also received several awards from IEEE conferences, including IEEE ICOIN Best Paper Award, in 2021; the IEEE Vehicular Technology Society (VTS) Seoul Chapter Awards, in 2019, 2021, and 2022; and the IEEE ICTC Best Paper Award, in 2022. He serves as an Editor and a Guest Editor for IEEE TRANSACTIONS ON VEHICULAR TECHNOLOGY, IEEE INTERNET OF THINGS JOURNAL, IEEE TRANSACTIONS ON MACHINE LEARNING IN COMMUNICATIONS AND NETWORKING, *IEEE Communications Standards Magazine*, *Computer Networks* (Elsevier), and *ICT Express* (Elsevier). He is a Distinguished Lecturer of the IEEE Communications Society (ComSoc) and the IEEE Systems Council.

SOYI JUNG (Member, IEEE) received the B.S., M.S., and Ph.D. degrees in electrical and computer engineering from Ajou University, Suwon, Republic of Korea, in 2013, 2015, and 2021, respectively.

She has been an Assistant Professor with the Department of Electrical of Computer Engineering, Ajou University, since September 2022. Before joining Ajou University, she was an Assistant Professor with Hallym University, Chuncheon, Republic of Korea, from 2021 to 2022; a Visiting Scholar with the Donald Bren School of Information and Computer Sciences, University of California, Irvine, CA, USA, from 2021 to 2022; a Research Professor with Korea University, Seoul, Republic of Korea, in 2021; and a Researcher with Korea Testing and Research (KTR) Institute, Gwacheon, Republic of Korea, from 2015 to 2016. Her current research interests include network optimization for autonomous vehicles communications, distributed system analysis, big-data processing platforms, and probabilistic access analysis. She was a recipient of the Best Paper Award by KICS, in 2015; the Young Women Researcher Award by WISET and KICS, in 2015; the Bronze Paper Award from IEEE Seoul Section Student Paper Contest, in 2018; the ICT Paper Contest Award by *Electronic Times*, in 2019; and the IEEE ICOIN Best Paper Award, in 2021.