A QoS Improving Downlink Scheduling Scheme for Slicing in 5G Radio Access Network (RAN)

Manoj Kumar Rana, Tommaso Pecorella *Senior Member, IEEE,* Bhaskar Sardar, Rama Rao Thipparaju *Senior Member, IEEE,* and Debashis Saha

***Abstract*—The 5G standard is aimed at supporting Quality of Service (QoS)-constrained traffic types, enabling new services to be reliably built into scenarios such as industrial automation and smart cities. The support comes via a strong emphasis on resource virtualization in the form of slices. Due to the strong QoS constraints of each slice, determining how to actually split the radio resources among different slices, while considering simultaneously the priority of slices, network efficiency, and each slice's target QoS, is very challenging. In this paper, we propose a radio resource scheduling scheme, designed on the basis of a strong theoretical analysis, to address the challenges. We formulate a Chance-constrained optimum resource allocation problem, which is then converted into a low complexity deterministic knapsack problem utilizing the concept of effective bandwidth. The performance analysis proves that our proposal is better in efficiency than the existing schemes, under different network conditions and QoS constraints. Results clearly show the effectiveness of our scheme in the considered 5G scenarios.**

***Index Terms*—5G, Network slicing, Resource management, Priority scheduling, Quality-of-service (QoS).**

I. Introduction

In the last decade, a novel technical evolution known as *virtualization* has deeply influenced the modern cellular systems. Evidently today a new operator can hardly deploy a full, greenfield nation-wide infrastructure. Instead, it is a common practice to create virtual operators (aka *tenants*) using the physical infrastructure of one or more telecommunication pipe providers. Virtualization allows a greater flexibility in the core network, enabling sharing of the core network resources among different tenants. This step in the virtualization process, fully embraced in the 5G architecture, is called network *slicing* [1]. In this approach, the network resources are not anymore owned by the tenants. Instead, it is seen as a resource pool, administrated by a manager super-party, for a set of tenants. The allocation of slices to the tenants is dynamic, and each tenant is characterized by its own target QoS. Depending on the tenant's goal and service model, different network slicing models are possible in the 5G architecture [2]. It is interesting to observe that due to non-constant traffic volume of each tenant, the amount of resources allocated to each slice must also be dynamically adjusted, taking into consideration simultaneously the network profit and each tenant's target QoS. As a consequence, slicing at inter-cell interference coordination level or at packet scheduling level [3] is more suitable in this case.

One of the main problems in packet scheduling is it's inability to fulfil the required end-to-end QoS constraints, mainly because the delays introduced in the Internet are random and not predictable. This is in contrast with the classical approaches, which consider mainly the user movements [4] or the wireless link variability [5].

The current state of the art makes use of a common packet scheduling function for all the slices in a cell mostly, without enforcing slice-specific treatment much. Few existing schemes use fixed prioritization to deal with heterogeneous slice types (see Section II). But the Internet induced random delay and load variation result into poor services for the slices [1], [3]. It is obvious that fixed prioritization will always decrease the performance of a higher QoS-driven application while running under a low priority slice (e.g., a live video streaming running in a simple video slice). On the other hand, conventional dedicated resource reservation policy ensures minimum resources for every slice and minimizes the effect of random delay and load variation. But, the 3rd Generation Partnership Project (3GPP) is of the opinion that 5G spectral efficiency can hardly be achieved this way. Flexible resource block configuration may improve spectral efficiency; but it is a very costly technique in terms of power consumption, delay and traffic overhead in the core.

The scheme proposed in this paper considers multiple slices with heterogeneous target QoS demands. The goal of our proposed scheme is set to provide optimum performance of each slice while strictly maintaining QoS constraints in the presence of traffic dynamics. We have realized the goal by formulating a Chance-constrained optimum resource allocation problem [6]. Then we have extended the concept of effective bandwidth [7] to convert the problem into a deterministic knapsack problem. Our proposed scheme ensures strict QoS demand fulfilment of every slice even in high load situation, allowing each slice to accommodate a large number of users. With the proposed method, Radio Access Network (RAN) slicing becomes possible without employing any costly static allocation process. In our solution, the amount of resources

M. K. Rana is with the Department of Computing Technologies, Faculty of Engineering and Technology, SRM Institute of Science and Technology, Kattankulathur, Tamil Nadu 603203, INDIA, (e-mail: manoj24.rana@gmail.com).

T. Pecorella is with the Information Engineering Department, Università di Firenze, Firenze, ITALY, (e-mail: tommaso.pecorella@unifi.it).

B. Sardar is with the Department of Information Technology, Jadavpur University, Kolkata, INDIA, (e-mail: bhaskargit@yahoo.co.in).

T. Rama Rao is with the Department of Electronics & Communication Engineering, Faculty of Engineering and Technology, SRM Institute of Science and Technology, Kattankulathur, Tamil Nadu 603203, INDIA, (e-mail: ramaraotr@gmail.com).

D. Saha is with the Management Information Systems Group, Indian Institute of Management Calcutta, Kolkata, INDIA, (e-mail: ds@iimcal.ac.in).

Manuscript received xx, xxxx; revised xx, xxxx.

to be reserved for each slice is dynamic and is calculated to satisfy the objective of optimum resources allocation with slice priority in consideration.

We have studied the performance of our proposed scheme in 5G-LENA [8], a New Radio (NR) network simulator, designed as a pluggable module to network simulator 3 (ns-3) project. We have compared our proposed scheme with the most relevant extant schemes which can handle heterogeneous QoS demanding slices, such as QoS aware schemes like Frame level scheduler (FLS) [9], Intelligent resource scheduling strategy (iRSS) [10] and Configuration-based assignment and packing (CBAP) algorithm [11]. Results clearly demonstrate that our scheme improves significantly the performance of every slice in terms of goodput and packet loss ratio (PLR) with respect to the existing schemes under different network conditions. Here, PLR represents the ratio of the number of lost packets to the total number of sent packets. A packet is may be either damaged due to a bad network condition or expired due to crossing its delay or inter-packet delay threshold, both of which are considered lost. We have shown that the priority of a slice can be controlled according to the current load and QoS demand of different applications, running under the slice. We have also demonstrated that the performance of a high priority slice is not affected even when the load of a low priority slice is increased, ensuring that slices prioritization is fulfilled always.

The main contributions of this paper are summarized as follows.

- Unlike existing works, we have considered both end-to-end and inter-packet delay as QoS parameters. Our scheduler guarantees to schedule a packet before any of these QoS parameters expire resulting in improved PLR and throughput compared to the existing schemes.
- We have proposed a novel dynamic prioritization scheme to control the service level of slices at each scheduling interval. This will allow the tenant operators to implement fine-grained policies for their designated slices. Also, to avoid starvation, traffic with varying priorities is scheduled in a non-sequential manner. Very few schemes of this kind exist, but they fail to maintain the service requirement for individual slices.
- We have used dissimilar packet sizes, Resource Block (RB) structures, and Next Generation NodeB (gNB)) configurations in our problem formulation. Accordingly, we have simulated a 5G heterogeneous network with overlapped macro and micro cells (operating at mm-wave bands). The proposed scheme based on this typical 5G network scenario is very useful for 4G network operators who want to deploy 5G networks in an incremental fashion.
- We have converted a complex NP-hard scheduling problem to a simple, deterministic knapsack problem that can be solved using a linear time complexity-based approach. On the other hand, the machine learning and reinforcement learning-based approaches used in existing literature are avoided due to their slow convergence and lack of support for specific QoS constraints.

TABLE I: Classification of existing radio resource scheduling schemes

Sequential	Ex-PF [12], M-LWDF [12], QTFDPS [13], PARS [14], TVS [15], RSESS [16], OCRDF [17], [18]
Semi-sequential	JRNSPSA [19], PPM [20], DARMA [21]
Non-sequential	FLS [9], iRSS [10], CBAP [11], JSBO [22], DOSSSVN [23], DSS [24], **Our Proposed Scheme**

The rest of the paper is organized as follows. In Section II, current state of the art on resource scheduling is presented. Section III illustrates the system model of our proposed scheme. In Section IV, our proposed scheme is designed. Performance analysis by simulation results is reported in Section V, and finally, conclusions are drawn in Section VI.

II. Related works

In the RAN slicing model, the virtual RAN infrastructure provider arranges all spectrum resources into a pool of carriers (ranging from 1.4 MHz to 400 MHz). The carriers are distributed for all the RANs on tenant-basis [25], where slicing is adopted at the spectrum planning level. The tenant-basis distribution assigns distinct set of carriers to every tenant across cells. In this way, tenant-specific function (e.g., scheduling algorithm) and policy can be applied. A complex resource scheduling scheme is needed to achieve tenant-specific QoS goals, especially for what concerns traffic prioritization.

Conventional scheduling schemes [12], working for single slice, are irrelevant here. These schemes include Blind Equal Throughput (BET), Proportional fair (PF), Exp rule, Log rule, Fair allocation high throughput (FAHT), etc. These approaches are exclusively designed for similar QoS demanding flows. As a consequence, they cannot be directly used as multi-slice schedulers in the 5G network.

In a network slicing scenario, a tenant will request one or more slices, and each slice will carry one or more traffic flows having dissimilar QoS requirements, e.g., VoIP, enhanced mobile broadband (eMBB), massive machine type communication (mMTC), etc. Depending upon the execution process of different priority slices, the existing radio resource scheduling schemes, as shown in Table I, can be categorized as: *sequential*, *semi-sequential* and *non-sequential*.

In the *sequential* approach, lower priority slice is scheduled after scheduling the higher priority one with the combination of the above-mentioned basic schemes. The Exponential PF (Ex-PF) and Modified largest weighted delay first (M-LWDF) schemes deal with Real-time (RT) and Non-Real-Time (NRT) flows differently [12]. These schemes schedule the first one by PF and the second one by a modified PF scheme, considering Head-of-line (HOL) delay and PLR as input parameters. Similarly, the QoS-oriented time and frequency domain packet scheduler (QTFDPS), proposed in [13], uses BET and PF for the NRT and RT flows, respectively. Another approach, Pricing-aware resource scheduling (PARS), proposed in [14], uses two advanced schemes: FLS [9] and M-LWDF [12] to

increase operator's revenue in case of heterogeneous subscription level (in terms of price) of users within the same slice.

In the above-mentioned schemes, scheduling is only considered between RT and NRT, but in 5G, NRT flows are expected to be low in number. Two-level Virtual Scheduler (TVS), proposed in [15], and RAN slicing with EDF slice scheduling (RSESS), proposed in [16], consider multiple RT slices, but resource scheduling is still sequential.

Another two novel approaches, the Optimal coverage and rate demand fulfillment (OCRDF) scheme and its extension, proposed in [17] and [18], respectively, target optimal allocation of resources among slices while their data rate demands are heterogeneous with each other. Although the authors take into account UE's location and channel condition to determine the data rate demand of a slice, the variation of demands at different time intervals due to uneven arrivals of packets even for the same set of UEs is not considered. Moreover, the absence of an appropriate strategy to handle QoS parameters, such as delay and inter-packet delay, sometimes results in scheduling precious resources to expired packets, and even in the worst situations, packets that are going to expire soon may be delayed while those that have sufficient time to expire are scheduled beforehand. Above all, these two schemes are also sequential in nature, as the lower priority slice is scheduled after the higher priority one.

The *sequential* approaches are efficient in under-loaded situations, whereas in moderate or overloaded conditions, the flows of low priority slices compete with each other, thereby getting very poor service; so these approaches are not suitable for 5G networks.

The *semi-sequential* approaches usually work sequentially but for some specific network conditions, where they work non-sequentially. In [19], a joint RT and NRT sliced packet scheduling and RB allocation (JRNSPSA) scheme use Exp rule [12] as the base scheduling strategy. When RT and NRT both enjoy medium or good channel condition, the scheme works sequentially. However, a NRT slice with good channel condition is scheduled before a RT slice with very bad channel condition. In [20], the Packet prediction mechanism (PPM) scheme is proposed, where a user is going to cross its PLR threshold, a packet of another user having PLR lower than its threshold can be delayed. Although it is a *sequential* approach, the delaying process always gives some room for the low priority slices over the higher ones and so, it acts as a *non-sequential* approach. Another approach, delay-aware resource management algorithm (DARMA), proposed in [21], preserves some RBs for RTs before scheduling. During scheduling of the RT slice, if some of the packets are very close to expiration, it takes a fraction of RBs from NRT slice. The conservation of RBs for NRTs makes this scheme *non-sequential*. However, high density of RTs does not leave any RBs for NRTs. PPM may be the best as it executes non-sequentially most of the time and improves in PLR but cannot satisfy high requirement of throughput and inter-packet delay constraints.

One of the basic *non-sequential* approaches is the two-level downlink scheduler for RT multimedia services, FLS is proposed in [9]. It is based on discrete time linear control theory. It significantly improves the throughput, PLR, fairness, and Quality of Experience (QoE). Though it considers end-to-end delay, it ignores inter-packet delay, and does not guarantee the packet loss of higher priority slices while serving lower priority slices. A slice-based non-sequential approach, iRSS is proposed in [10] where optimum resource utilization is done by intelligent prediction of current resource block allocation from the past statistics through collaborative learning. Although the deviation of their prediction from the actual allocation is minimized, they do not consider heterogeneous QoS.

A delay-sensitive cell-level approach [11] proposed the CBAP algorithm which targets to minimize the number of scheduled, but not-served QoS flows by configuring the frame with dynamic-sized RBs for multiple QoS flows. A similar non-sequential approach is also proposed in [26], where mini-slot-based resource allocation is used to accommodate upcoming URLLC packets inside a large eMBB frame. The main goal is to maximize the data rate of eMBB users while satisfying URLLC's delay constraint. However, they do not take into account the heterogeneous QoS demands of different eMBB flows. As they are designed for single cell only and consider similar kind of QoS flows, it can hardly cope up in heterogeneous 5G network slicing scenario.

A Joint scheduling and beam-forming optimization (JSBO) in Software Defined Network (SDN)-based virtual wireless environments, containing massive number of IoT devices, is proposed in [22] to minimize power consumption. A non-cellular Delay optimal stochastic scheduling scheme for vehicular networks (DOSSSVN) is proposed in [23] which targets to minimize the delay of the vehicular network applications. They have considered high fluctuation of vehicle arrival rate and optimize the allocation with respect to delay, but this strategy does not suit heterogeneous 5G network slicing scenario.

The dynamic slicing-based scheme (DSS), proposed in [24], only keeps the minimum resource reservation constraint, allocating resources non-sequentially among different slices. It uses a modified PF scheme considering only fairness among slices, and channel condition but it does not consider QoS constraints. Due to non-sequential nature, it can perform better than the other two approaches but insensitivity to QoS makes it unsuitable to manage heterogeneous 5G network slices effectively.

From the above discussion, it becomes clear that none of the existing schemes is perfectly suitable for handling heterogeneous 5G network slicing scenarios. To address this gap, we propose a *non-sequential*, QoS-aware scheduling scheme that ensures effective prioritization of slices. Our scheme comes close to the FLS [9], iRSS [10], and CBAP [11] schemes. But it differs from each of them in terms of QoS parameter (e.g., inter-packet delay), dynamic prioritization of slices, low complexity, cost effectiveness, and flexibility in service provisioning of slices.

III. System Model

The system model illustrates the network topology, traffic model and the statistical distribution used, followed by the

objective function. They are presented in the following.

A. Network Topology

In this paper, we have considered multiple network slices in Network Function Virtualization (NFV)/ SDN enabled integrated mobile network. The mobile network may contain multiple Base Stations (BSs) with different configuration of resource blocks (RBs) in terms of their bandwidth and time span [11]. The spectrum resources of multiple BSs in terms of RBs are aggregated into a single resource pool. The data efficiency of a particular RB with respect to a User Equipment (UE) is a function of received power, intra and inter-cell interference, Additive White Gaussian noise (AWGN) and the size of the RB [11]. Let us assume that there exist some Distributed Radio Resource Management (DRRM) modules, integrated into the nearby Data Center (DC), to collect the scheduling and traffic information like delay, inter-packet delay, packet arrival rate, received power, etc. from the RRM module of individual BSs.

B. Traffic Model

We have mainly focused on the downlink traffic. A *slice* may contain a group of traffic flows requiring different target QoS demands. Let us denote the set of *slices* as $\mathcal{C} = \{1, .., i, ..., C\}$ and by $U^{(i)}$, $i \in \mathcal{C}$, we denote the subset of packets belonging to *slice* i. If U is the set of all packets under the RAN, we can write:

$$U = \bigcup_{i=1}^{C} U^{(i)} \tag{1}$$

We have characterized the QoS of each traffic flow by end-to-end delay and inter-packet delay parameters. Threshold values for these QoS parameters for every QoS class are defined in [12]. In order not to violate the end-to-end delay and the inter-packet delay constraints, the scheduler must serve its queued packets within an *expiration time threshold*, denoted as Δ. So, a packet will be expired at $(T + \Delta)$, where T is the current time. Our system model is dynamic in the sense that a scheduler is implemented at every time slot, i.e., it implicitly takes care of the status of active slices, including queued-up packets, packet delays, and so on. Each packet within the subset $U^{(i)}$ may have different value of Δ due to previous delays in the network.

$$\Delta = \min_{\forall q} \left\lceil \left(t_{thr}^{(q)} - t^{(q)} \right) \right\rceil \tag{2}$$

where $t^{(q)}$ and $t_{thr}^{(q)}$ are respectively the computed value and the threshold value of the QoS parameter, q.

The end-to-end delay contains two components: 1) the delay between the remote server and the 5G gateway (User Plane Function (UPF)), denoted as $d_{Server,UPF}$, and 2) the delay between the UPF and the UE, denoted as $d_{UPF,UE}$. Consequently, the following equation holds.

$$t^{(Pkt-Delay)} = d_{Server,UPF} + d_{UPF,UE} \tag{3}$$

The current delay-based resource scheduling schemes [9] consider only $d_{UPF,UE}$ because 3GPP has specified the delay threshold between UPF and UE. But, it can not be considered as the end-to-end delay threshold.

TABLE II: List of notations

Symbol	Definition
U	Set of all downlink packets
$U^{(i)}$	Set of all downlink packet of the i^{th} slice, $U^i \subseteq U$
T	Current time (in DTI)
Δ	Expiration time threshold (in DTI) of a packet, $0 \leqslant \Delta \leqslant M$
M	Maximum value of Δ
$U_{T+\Delta}^{(i)}$	Set of all downlinks packets of i^{th} slice having expiration time threshold Δ, $U_{T+\Delta}^{i} \subseteq U^i$
μ	Mean of a random variable
S	Total number of states in the Markov chain of MTC traffic
$\mathbf{P}$	State transition probability matrix in MTC
$\overline{L_i}$	Mean sojourn time in state i
$\overline{W}$	Mean traffic arrival rate of MTC
R	Total number of Resource Blocks
$\mathcal{R}$	Total capacity of the resource pool, $\mathcal{R} > 1$
$d_j^{(i)}$	Transmission need of the i^{th} sliceat at j^{th} DTI
$\varphi_{j,k}^{(i)}$	Average number of packets per subset calculated upto k^{th} subset at j^{th} DTI for i^{th} slice
$\beta_{j,k}^{(i)}$	Number of new packets arrived up to k^{th} subset at j^{th} DTI for i^{th} slice
$\overline{W}_i$	Mean arrival rate at state, i of MTC traffic
δ_i	Priority value of the i^{th} packet
y_i	Binary decision variable (Takes value 1 if i^{th} packet is selected, otherwise takes 0)
b_i	Required data capacity of i^{th} packet
H_{ij}	Data efficiency of i^{th} packet over j^{th} RB
$\mathcal{Z}_i$	Amount of resource, allocated to i^{th} packet
ψ_i	Allocated resource minus the capacity of i^{th} packet
p	Overflow probability
γ	Overflow probability while Sample Avg. Approx. is used
ξ	Probability of $\hat{S}_{\gamma}^{\mathcal{N}} > S_p^*$
$\mathcal{N}$	No. of Monte Carlo samples of the capacity constraint of the objective function

One-way delay measurement is very challenging due to lack of cooperation between the remote server/end node and the 5G network [27]. We have assumed that the UPF can monitor the delay between itself and the remote server by resorting to *ping*, *traceroute*, or any other means, e.g., by resorting to Real Time Protocol features, IPv6 Performance and Diagnostic Metrics (PDM) [28], etc. The delay between the UPF and the UE can be measured by using a probing technique [27] where every fragment of a test packet is time stamped at both the UPF and the UE. Then $d_{UPF,UE}$ can be computed as:

$$d_{UPF,UE} = |max_k(TS_{UPF,k}) - max_k(TS_{UE,k})| \tag{4}$$

where $TS_{UPF/UE,k}$ is the time stamp of k^{th} fragment at the UPF/UE. As the fragments, generated at the UPF, must be reconstructed at the UE to get back the original datagram, independent of the routing paths, the maximum values of the time stamps are used in (4).

The computed value of inter-packet delay in the downlink direction is the difference between the current time and the last time packet was received by the UE.

To group packets having same expiration time, we have defined a new subset, denoted as $U_{T+\Delta}^{(i)}$, which contains packets having same Δ. Thus, the i^{th} packet group can be

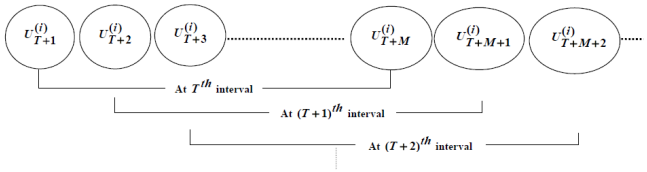

Fig. 1: Expiration time illustration for different subset

expressed as:

$$U^{(i)} = \bigcup_{\Delta=1}^{M} U^{(i)}_{T+\Delta} \tag{5}$$

where Δ is a random variable which follows certain distribution with values ranging from 0 to the maximum, M. The value of M may have a very high value if the running application can sustain a very high delay or inter-packet delay threshold values. However, we assume that the threshold value is restricted within a certain limit for the selected flows of the slices. For a packet, Δ is less than or equal to 0, then it is already expired and it will be discarded.

The illustration of valid subsets in different time intervals is shown in Fig. 1. In the $(T+1)^{\text{th}}$ time interval, the packets containing in $U^{(i)}_{T+1}$ is expired as $\Delta = 0$ and hence it is deleted. A new subset, $U^{(i)}_{T+M+1}$, is included as total M number of subsets are valid in every time interval.

C. Statistical Distributions Used

The traffic arrival process is often modeled as a simple Poisson distribution. But arrival of traffic like VoIP (e.g., Skype voice call, Google Talk, QQ chat), HTTP (e.g., Google searching, Facebook) can be modeled as a modified Poisson distribution [29] or, non-homogeneous Poisson process [30], while inter-arrival time of video streaming (e.g., YouTube, Netflix, Hotstar live) can be modeled by Pareto distribution [31]. The mMTC traffic can be modeled using a Semi-Markov Model (SMM) [32].

A modified Poisson distribution is a linear transformation of Poisson distribution as defined in [29]. It can be specified by three parameters (a, b, λ). If X is a Poisson random variable with expected rate of occurrences as λ, the modified Poisson random variable Y is defined as:

$$Y = aX + b \tag{6}$$

The Probability Density Function (PDF) of Y can be derived as:

$$\Pr\{Y = y\} = \frac{\lambda^{\frac{y-b}{a}} e^{-\lambda}}{\frac{y-b}{a}!} \tag{7}$$

The mean (μ) of Y can be derived using (7) as follows:

$$\mu = \mathbb{E}[Y] = a\lambda + b \tag{8}$$

If Z is a random variable of inter-arrival time following Pareto distribution, the probability distribution can be given by:

$$\Pr\{Z = z\} = \begin{cases} \frac{\alpha z_m^{\alpha}}{z^{\alpha+1}} & \text{for } z \geqslant z_m \\ 0 & \text{for } z < z_m \end{cases} \tag{9}$$

where z_m is the minimum possible value of Z, and α is a positive shape parameter. The mean of Z can be given as:

$$\mu = \frac{\alpha z_m}{\alpha - 1}, \alpha > 1 \tag{10}$$

The mMTC has different traffic patterns. These patterns can be integrated into a Markov structure with S number of states [32]. Let $\mathbf{P}$ be the state transition matrix, and $p_{i,j}$ be the transition probability from state i to j. The stationary state probabilities of the embedded Markov chain $\mathbf{\Pi}^{(e)}$ can be obtained by the following eigenvalue problem:

$$\mathbf{\Pi}^{(e)} = \mathbf{\Pi}^{(e)}\mathbf{P} \quad \text{where} \sum_{i=1}^{S} \Pi_i^{(e)} = 1 \tag{11}$$

If the mean sojourn time in state i is $\overline{L_i}$, the actual state probabilities can be computed by using the following formula:

$$\Pi_i = \frac{\Pi_i^{(e)}\overline{L_i}}{\sum_{j=1}^{S} \Pi_j^{(e)}\overline{L_j}} \tag{12}$$

We assume that the arrival rate of packets in each state is a fixed process. If $\overline{W_i}$ is the arrival rate at state i, the mean arrival rate $\overline{W}$ can be calculated as follows:

$$\overline{W} = \sum_{i=1}^{S} \Pi_i \overline{W_i} \tag{13}$$

D. Objective Function

The objective of our proposed scheme is to schedule the packets of multiple slices in a non-sequential manner such that every slice gets optimum performance while strictly maintaining the QoS constraints. We have introduced some parameters, like H_{ij}, denoting the data efficiency of i^{th} packet in j^{th} RB, δ_i denoting the priority of i^{th} packet, x_{ij} denoting a binary decision variable whether j^{th} RB is selected for i^{th} packet or not, y_i denoting a binary decision variable whether i^{th} packet is selected for scheduling or not and b_i to denote the required data capacity of i^{th} packet. As given in [33], H_{ij} can be defined as follows:

$$H_{ij} = \theta_j \eta \log_2(1 + \frac{P_{ij}^{(rcv)}}{I_{ij}^{(int)} + P_{ij}^{(awgn)}}) \tag{14}$$

where θ_j denotes the size of the j^{th} RB, $\eta (0 \leqslant \eta \leqslant 1)$ is the attenuation factor accounting to implementation, and $p_{ij}^{(rcv)}$, $I_{ij}^{(int)}$ and $P_{ij}^{(awgn)}$ are representing received, intra-cell interference and AWGN power at the UE having i^{th} packet and located at the BS containing j^{th} RB. Hence the objective function can be written as follows:

$$\max \sum_{i=1}^{|U|} \delta_i y_i \tag{15}$$

subject to

$$\sum_{j=1}^{R} H_{ij} x_{ij} \geqslant b_i, \forall i, \tag{16}$$

$$x_{ij} \in \{0, 1\}, \tag{17}$$

$$y_i \in \{0, 1\} \tag{18}$$

The above problem is NP hard and the proof is given in Appendix A. In (16), to include a packet into our selection grid, b_i must be satisfied by one or more number of RBs from the resource pool. Let us assume that the amount of radio resources, allocated to satisfy i^{th} packet's capacity constraint minus its actual requirement be a random variable, ψ_i and the total capacity of the resource pool is $\mathcal{R}$ unit ($\mathcal{R} > 1$), taken as a shape parameter. Then the amount of resources allocated to the i^{th} packet becomes a random variable, denoted as $\mathcal{Z}_i$, which can be defined as follows:

$$\mathcal{Z}_i = \frac{b_i + \psi_i}{\sum_{j=1}^{R} H_{ij}} \mathcal{R} \tag{19}$$

Now the above optimization problem can be converted into a Chance-constrained problem [6]. According to the definition of Chance-constrained problem [6], the above problem can be approximated as:

$$S_p^* = \max \sum_{i=1}^{|U|} \delta_i y_i \tag{20}$$

subject to

$$Pr(\sum_{i=1}^{|U|} \mathcal{Z}_i y_i > \mathcal{R}) \leqslant p \tag{21}$$

$$y_i \in \{0, 1\} \tag{22}$$

where p is the probability of the occurrence that the total resource requirement of the selected packets exceeds the capacity of the resource pool and S_p^* is the optimal value of the objective function with parameter p. The above problem would give us the solution vector of the selected packets i.e., $\vec{Y} = y_1, \ldots, y_{|U|}$ with a risk factor of p which is considered as small as possible.

The priority vector $\vec{\delta} = \{\delta_1, \ldots, \delta_{|U|}\}$ can be modeled as an input to the objective function. Both static and dynamic prioritizations are possible. In case of static prioritization, a packet $k \in U$, can be scheduled before any packet of the subset $\hat{U} \subset U$ if the following inequality holds,

$$\delta_k > \sum_{w \in \hat{U}} \delta_w, k \notin \hat{U} \tag{23}$$

We have used the above inequality for all packets $k \in U_{T+1}^i$ and $\forall i \in \mathcal{C}$ i.e., the set of packets which will expire if not scheduled at current DTI. The dynamic prioritization is used to determine the priority of packets among slices. Instead of satisfying the above inequality, δ_k is set at a lower value such that a slice will also get some room for its packets to be scheduled before the packets of a higher priority slice. Note that, in this case δ_k depends on the objective function itself. Therefore, we can obtain δ_k for the packets of different slices in case of dynamic prioritization through simulation only.

IV. Proposed scheme

We have developed an efficient algorithm to compute the optimum allocation of available RBs to all the packets in Section IV-A. In the Section IV-B, we have discussed the minimum transmission need at the current DTI. The complexity of the proposed scheme is discussed in Section IV-C.

A. The Allocation Algorithm

The main idea is to convert the objective function, defined in (20), (21) and (22) into a deterministic knapsack problem by calculating the effective bandwidth [7] of the random variable $\mathcal{Z}_i$, given in (19). Let the effective bandwidth of the random variable $\mathcal{Z}$ is denoted as $\Gamma_p(\mathcal{Z})$ with overflow probability p. The standard effective bandwidth of $\mathcal{Z}$ is defined as follows [7]:

$$\Gamma_p(\mathcal{Z}) = \frac{\log E[p^{-\mathcal{Z}}]}{\log p^{-1}} \tag{24}$$

Proposition 1. *Let $\mathcal{Z}_1, \ldots, \mathcal{Z}_n$ be independent random variables and $\mathcal{Z} = \sum_i \mathcal{Z}_i$. Let $g > h$. If $\sum_i \Gamma_p(\mathcal{Z}_i) \leqslant h$, then $Pr[\mathcal{Z} \geqslant g] \leqslant p^{g-h}$.*

Proof. If $\sum_i \Gamma_p(\mathcal{Z}_i) \leqslant h$, then from the standard definition of effective bandwidth, given in (24) we can say that $\sum_i \log E[p^{-\mathcal{Z}_i}] \leqslant \log p^{-h} \Rightarrow \prod_i E[p^{-\mathcal{Z}_i}] \leqslant p^{-h}$.

Now, $Pr[\mathcal{Z} \geqslant g] = Pr[p^{-\mathcal{Z}} \geqslant p^{-g}] \leqslant p^g E[p^{-\mathcal{Z}}]$ (following Markov's inequality). Hence, $p^g E[p^{-\mathcal{Z}}] = p^g \prod_i E[p^{-\mathcal{Z}_i}] \leqslant p^{g-h}$ (from the above inequality). □

In the inequality, proved in Proposition 1, if we set $g = \mathcal{R}$, $h = g - 1$ and use the condition given below:

$$\Gamma_p(\mathcal{Z}_i y_i) \leqslant \mathcal{R} - 1 \tag{25}$$

we can get, $Pr[\sum_{i=1}^{|U|} \mathcal{Z}_i y_i > \mathcal{R}] \leqslant p$, where $\Gamma_p(\mathcal{R}) = \mathcal{R} - 1$. To satisfy the condition, $\Gamma_p(\mathcal{Z}_i y_i) \leqslant (\mathcal{R} - 1)$, we have considered sample average approximation approach [34]. Let $\mathcal{Z}^{(1)}, \ldots, \mathcal{Z}^{(\mathcal{N})}$ be $\mathcal{N}$ number of independent Monte Carlo samples of the random variable $\mathcal{Z}$. Let us denote the optimal value of the objective function with parameters γ and $\mathcal{N}$ as $\hat{S}_\gamma^{(\mathcal{N})}$, where γ is the overflow probability while sample average approximation approach is used [34]. For $\gamma = [0, 1]$, the problem can be redefined as:

$$\hat{S}_\gamma^{(\mathcal{N})} = max \sum_{i=1}^{|U|} \delta_i y_i \tag{26}$$

subject to

$$\frac{1}{\mathcal{N}} \sum_{k=1}^{\mathcal{N}} \mathcal{I} \left[\sum_{i=1}^{|U|} \mathcal{Z}_i^{(k)} y_i^{(k)} - (\mathcal{R} - 1) \right] \leqslant \gamma \tag{27}$$

$$y_i \in \{0, 1\} \tag{28}$$

where $\mathcal{I}[.]$ is the indicator function, such that $\mathcal{I}[.] = 1$ if $'.' > 0$ and $\mathcal{I}[.] = 0$ otherwise. Here, we have taken slightly more risk, i.e., $\gamma > p$. But, we have shown that the solution from this problem is no worse than the optimal. If the previous problem is denoted as $Prob_p^*$ and the current problem is denoted as $Prob_\gamma^{(\mathcal{N})}$, the following theorem proves that by taking a risk parameter $\gamma > p$ in our problem the optimal value, $\hat{S}_\gamma^{(\mathcal{N})}$, will be an upper bound to the true optimal, S_p^*, with the probability approaching 1 exponentially fast as $\mathcal{N}$ increases.

Theorem 1. *Let $\gamma > p$ and the $Prob_p^*$ has an optimal solution, then*

$$Pr(\hat{S}_\gamma^{(\mathcal{N})} \geqslant S_p^*) \geqslant 1 - \exp\{-2\mathcal{N}(\gamma - p)^2\} \tag{29}$$

Proof. The proof is given in [34]. □

If we take the value of $\mathcal{N}$ sufficiently large such that the value of the right-side of the inequality (29) becomes close to 1, it could be said that the solution of $Prob_{\gamma}^{(\mathcal{N})}$ is equivalent to solution of $Prob_{p}^{*}$. Taking a specified a confidence probability of $(1-\xi)$, we could determine the value of $\mathcal{N}$ in the following proposition.

Proposition 2. *The following choice of $\mathcal{N}$ ensures $\hat{S}_{\gamma}^{(\mathcal{N})} \geqslant S_{p}^{*}$ with a confidence probability of $(1-\xi)$, where $\gamma > p$.*

$$\mathcal{N} \geqslant \frac{1}{2(\gamma-p)^2} \log \frac{1}{\xi} \tag{30}$$

Proof. If the confidence probability is $(1-\xi)$, from Theorem 1, we can write:

$Pr(\hat{S}_{\gamma}^{(\mathcal{N})} \geqslant S_{p}^{*}) \geqslant 1 - \exp\{-2\mathcal{N}(\gamma-p)^2\} \geqslant 1-\xi$

$\xi \geqslant \exp\{-2\mathcal{N}(\gamma-p)^2\}$

$\mathcal{N} \geqslant \frac{1}{2(\gamma-p)^2} \log \frac{1}{\xi}$ □

Hence, we can determine the lower bound of $\mathcal{N}$, and by increasing it such the sample values satisfy the condition, given in (27), we can find an optimal solution for the decision vector. Assuming a considerable value of the confidence interval, $(1-\xi)$, the above procedure ensures that the condition in (25) will be satisfied as proved in Theorem 2.

Theorem 2. *Using the capacity constraint, given in* (27)*, we can prove that $Pr(\sum_{i=1}^{|U|} \mathcal{Z}_i y_i > b) \leqslant p$.*

Proof. Applying Proposition 1 in Theorem 1, we have proved the optimal solution in $Prob_{\gamma}^{(\mathcal{N})}$ is equivalent to the optimal solution obtained from $Prob_{p}^{*}$ with a confidence probability $(1-\xi)$. So, the effective bandwidth of these two problems are equal with a confidence of $(1-\xi)$. For $k \in \mathcal{N}$, if $\xi \ll 1$, we can write,

$$\begin{aligned} \sum_{i=1}^{|U|} \Gamma_p\left(\frac{\mathcal{Z}_i y_i}{\mathcal{R}-1}\right) &\approx \sum_{i=1}^{|U|} \Gamma_\gamma\left(\frac{\mathcal{Z}_i^{(k)} y_i}{\mathcal{R}-1}\right) \\ &= \sum_{i=1}^{|U|} \frac{\log(E[\gamma^{-\frac{\mathcal{Z}_i^{(k)} y_i}{\mathcal{R}-1}}])}{\log(\gamma^{-1})} \end{aligned} \tag{31}$$

By using Jensen's inequality, we can write,

$$\begin{aligned} \sum_{i=1}^{|U|} \frac{\log(E[\gamma^{-\frac{\mathcal{Z}_i^{(k)} y_i}{\mathcal{R}-1}}])}{\log(\gamma^{-1})} &\leqslant \sum_{i=1}^{|U|} \frac{E[\log(\gamma^{-\frac{\mathcal{Z}_i^{(k)} y_i}{\mathcal{R}-1}})]}{\log(\gamma^{-1})} \\ &= \sum_{i=1}^{|U|} \frac{E[\mathcal{Z}_i^{(k)} y_i]}{(\mathcal{R}-1)} \end{aligned} \tag{32}$$

Now, the expected value of $\mathcal{Z}_i^{(k)} y_i$ over all the samples of $\mathcal{N}$ is at most $(\mathcal{R}-1)$, as $\gamma \ll 1$. So, we can write,

$$\sum_{i=1}^{|U|} \Gamma_p(\mathcal{Z}_i y_i) \leqslant \mathcal{R}-1 \tag{33}$$

According to the Proposition 1, we can write, $Pr(\sum_{i=1}^{|U|} \mathcal{Z}_i y_i > \mathcal{R}) \leqslant p$ □

Our problem can now be converted to a $0/1$ Knapsack problem:

$$S_{p}^{*} = \max \sum_{i=1}^{|U|} \delta_i y_i \tag{34}$$

subject to

$$\sum_{i=1}^{|U|} \Gamma_p(\mathcal{Z}_i) y_i \leqslant \mathcal{R}-1 \tag{35}$$

$$y_i \in \{0,1\} \tag{36}$$

To solve the above problem, we may use dynamic programming approach. But, due to pseudo polynomial time complexity of this kind of approach, we shall use a modified Linear Programming relaxation-based approach as given in [35], which has the time complexity $O(|U|)$.

B. Determination of Transmission Need

The transmission need, denoted as d_T, is the minimum number of packets to be selected at current DTI from the set U, such that no packet of this set will be dropped in future DTIs due to QoS constraints. So, we can write,

$$d_T = \sum_{i=1}^{C} d_T^{(i)} \tag{37}$$

where $d_T^{(i)}$ is the transmission need of the i^{th} *slice*.

Let $\varphi_{j,k}^{(i)}, j < k \leqslant j+M$ be the partial average of packets, i.e., the average number of packets per subset calculated up to k^{th} subset at j^{th} DTI for i^{th} slice. $\varphi_{j,k}^{(i)}$ can be expressed as:

$$\varphi_{j,k}^{(i)} = \frac{1}{k-j} \sum_{n=j+1}^{k} |U_n^{(i)}| \tag{38}$$

As the arrival process of packets in different subsets are dynamic with respect to time (DTI), the derivation of $d_T^{(i)}$ is complex. So we divide the derivation in two consecutive steps.

1) Step 1: Here, we have assumed no arrival of packets in future DTIs. Now, if all the packets of the subset $U^{(i)}$ are scheduled, i.e., $d_T^{(i)} = \left|U^{(i)}\right|$, our scheme would become *sequential*, i.e., higher priority slice will be scheduled fully before going for lower priority slice. The novel purpose of *non-sequential* scheduling cannot be achieved by this way.

As $\left|U_{T+1}^{(i)}\right|$ is the number of packets that will expire at $(T+1)^{\text{th}}$ DTI, they must be scheduled in T^{th} DTI. So, the following inequality must hold:

$$d_T^{(i)} \geqslant \left|U_{T+1}^{(i)}\right| \tag{39}$$

To achieve the goal of minimum packet selection, we could select the average over all the subsets of the set $U^{(i)}$, i.e., $\varphi_{T,T+M}^{(i)}$. However, this may be less than $\left|U_{T+1}^{(i)}\right|$. Hence, it would lead to packet loss at the current DTI. To avoid such losses, we can select the maximum between $\varphi_{T,T+M}^{(i)}$ and $\left|U_{T+1}^{(i)}\right|$. Now, let us assume the following inequality,

$$\varphi_{T,T+2}^{(i)} > \max\left(\left|U_{T+1}^{(i)}\right|, \varphi_{T,T+M}^{(i)}\right) \tag{40}$$

i.e., the partial average up to $(T+2)^{\text{th}}$ subset is greater than what we have selected in the current DTI. Hence, it is obvious that the selection at $(T+1)^{\text{th}}$ DTI will be greater than the selection at T^{th} DTI. Thus, if the value of $\varphi^{(i)}_{T,T+2}$ is very high, we may not be able to accommodate all the packets due to shortage of resources. Thus, the selection may increase in future DTIs, like at $(T+12)$, $(T+3)$, and so on. In this case, to make our scheme smoother, a bigger allocation can be started from the current DTI.

The above selection becomes invalid because of different values of the partial averages, $\varphi^{(i)}_{T,k}$, $T < k \leqslant T+M$. As the distribution of packets in different subset is dynamic, $\left|U^{(i)}_j\right|$, $T < j \leqslant T+M$, is also different. Hence, it is obvious that the partial average will be different and may be greater than the overall average. As a consequence, choosing the maximum partial average would be a better option than the previous choice. We need to prove that there is no packet loss in future DTIs using this selection process. This proof considers dynamic arrival of packets with respect to time.

Lemma 1. *If transmission need is the maximum partial average, then for $k \geqslant T$,*

$$\sum_{l=T}^{k} d^{(i)}_l \geqslant \sum_{l=T+1}^{k+1} \left|U^{(i)}_l\right| + \sum_{l=T+1}^{k} \beta^{(i)}_{l,k+1} \tag{41}$$

where $\beta^{(i)}_{m,n}$ is the number of new packets arrived up to n^{th} subset at m^{th} DTI for i^{th} slice. $\beta^{(i)}_{m,n} = 0$ if $m \geqslant n$.

Proof. At l^{th} DTI, $d^{(i)}_l$ denoting the maximum partial average, is always greater than or equal to any partial average, evaluated at that DTI. Hence, the following inequality must hold:

$$d^{(i)}_l \geqslant \varphi^{(i)}_{l,k+1}, \qquad \text{for } l \leqslant k \tag{42}$$

For $l = T$, $l = T+1$ and $l = T+2$, $\varphi^{(i)}_{l,k+1}$ can be written as follows.

$$\varphi^{(i)}_{T,k+1} = \frac{\sum_{l=T+1}^{k+1} |U^{(i)}_l|}{k-T+1} \tag{43}$$

$$\varphi^{(i)}_{T+1,k+1} = \frac{\sum_{l=T+1}^{k+1} |U^{(i)}_l| + \beta^{(i)}_{T+1,k+1} - d^{(i)}_T}{k-T} \tag{44}$$

$$\varphi^{(i)}_{T+2,k+1} = \frac{1}{k-T-1}\left(\sum_{l=T+1}^{k+1} |U^{(i)}_l| + \beta^{(i)}_{T+1,k+1} + \beta^{(i)}_{T+2,k+1} - d^{(i)}_T - d^{(i)}_{T+1}\right) \tag{45}$$

Deriving up to k^{th} DTI, we can write,

$$\varphi^{(i)}_{k,k+1} = \sum_{l=T+1}^{k+1} |U^{(i)}_l| + \sum_{l=T+1}^{k} \beta^{(i)}_{l,k+1} - \sum_{l=T}^{k-1} d^{(i)}_l \tag{46}$$

By substituting l with k in (42) and using (46), it can be proved that

$$\sum_{l=T}^{k} d^{(i)}_l \geqslant \sum_{l=T+1}^{k+1} |U^{(i)}_l| + \sum_{l=T+1}^{k} \beta^{(i)}_{l,k+1} \tag{47}$$

□

2) Step 2: In this step, we consider dynamic arrival of packets and their distribution in different subsets. It is proved in the above Lemma 1 that the total number of packets scheduled up to k^{th} DTI, where $k \geqslant T$, is always greater than or, equal to the total number of packets that must be scheduled within k^{th} DTI before their expiration. So, it is proved that maximum partial average can be a suitable parameter as it fulfils the goal of not dropping any packet in future.

Lemma 2. *If transmission need is the maximum partial average and m^{th} indexed partial average is selected at current DTI T ($m > T$), then $d^{(i)}_j \geqslant d^{(i)}_l$, where $T \leqslant l < j \leqslant m$.*

Proof. Let us assume that at l^{th} DTI, s^{th} indexed partial average is selected, where $s > l$. Now, at $(l+1)^{\text{th}}$ DTI, the new selection metric is always greater than or equal to any partial average, computed at that DTI. The s^{th} indexed partial average at $(l+1)^{\text{th}}$ DTI exist if $s > (l+1)$ and it can be computed as follows:

$$\varphi^{(i)}_{l+1,s} = d^{(i)}_l + \frac{\beta^{(i)}_{l+1,s}}{s-l-1} \tag{48}$$

The following inequality must hold.

$$d^{(i)}_{l+1} \geqslant \varphi^{(i)}_{l+1,s} \tag{49}$$

As $s > (l+1)$ and $\beta^{(i)}_{l+1,s} \geqslant 0$, using (48) and (49), we can write $d^{(i)}_{l+1} \geqslant d^{(i)}_l$ where $(l+1) \leqslant m$. Similarly, it can also be proved that $d^{(i)}_{l+2} \geqslant d^{(i)}_{l+1}$ when $(l+2) \leqslant m$ and so on. As $j > l$, it is now proved that

$$d^{(i)}_j \geqslant d^{(i)}_l \tag{50}$$

where $T \leqslant l < j \leqslant m$. □

If we select the above metric where m^{th} indexed partial average is maximum at T^{th} DTI, it has been proved in Lemma 2 that the function $d^{(i)}_j$ becomes a non-decreasing function for $T \leqslant j < m$. The non-decreasing nature of the function occurs because of the future arrival of packets up to m^{th} subset, i.e., $\sum_{l=T+1}^{m-1} \beta^{(i)}_{l,m}$. Hence, serving equally these future packets along with the maximum partial average in each DTI can make fairer scheduling than only maximum partial average selection. To determine average number of packets that will arrive in future, we use a prediction mechanism using the basic probability distributions as given in Section III-C. The mean number of packets that will arrive between T^{th} and $(k-1)^{\text{th}}$ DTIs and will be included up to k^{th} subset will be denoted as $f(T,k)$. Accordingly, $d^{(i)}_T$ can be written as follows:

$$d^{(i)}_T = \max_{k=T+1}^{T+M}\left(\varphi^{(i)}_{T,k} + \frac{f(T,k)}{k-T}\right) \tag{51}$$

In Appendix B, we have determined the value of $f(T,k)$ on the basis of the arrival process of the packet in the i^{th} slice, which is tied to the kind of traffic the slice is carrying.

According to (57), (58) and (59), we can rewrite $\frac{f(T,k)}{k-T}$ for voice, video, and MTC applications respectively as:

$$\frac{f\left(T,k\right)}{k-T}=\begin{cases}\frac{n\mu(k-T+1)}{2M} & \text{for voice traffic}\\ \frac{(k-T+1)}{2\mu M} & \text{for video traffic}\\ \frac{(k-T+1)\overline{W}}{2M} & \text{for MTC traffic}\end{cases} \tag{52}$$

Note that this process can be extended to other traffic types too.

In Proposition 1, provided in Appendix C, we have demonstrated that the transmission need parameter calculated using (51) is a decreasing function with respect to time under the assumption that the same number of packets are included in each subset in each DTI. The case of dynamic traffic arrival is illustrated by simulation in Section V.

C. Complexity Analysis

In our proposed scheme, we have suggested to use the algorithm, given in [35], to solve the basic $0/1$ knapsack problem whose time complexity is $O(|U|)$. The transmission need computation as given in Section IV-B is taken as input to this problem. The per slice transmission need computation, given in (51). The number of logical operations is M times the number of computations to determine the per-slice metric, $f(T,k)$. Any object-oriented design will follow a recursive process to compute it. First, We fetch the size of a subset. Then the number of computations needed to calculate the partial average, $\varphi_{T,k}^{(i)}$, is given as follows:

$$\#\text{computation}=\frac{r\left(k-T-1\right)+s}{k-T}+v \tag{53}$$

where r is the partial average up to $(k-1)^{\text{th}}$ subset, s is the size of the k^{th} subset and v is the parameter as provided in (52). The computation of r and s can be performed when the queue is updated, and can be neglected in the complexity analysis. The above equation contains 11 arithmetic operation and 2 fetching operations. So, the selection metric determination requires $13\times M$ number of operations in total which equals to $O\left(M\right)$. The physical allocation of RBs takes $O(\mathcal{R}^2N)$ times [12] where $\mathcal{R}$ and N are total number of RBs and total number of users respectively. Hence, the total time complexity becomes $O(M+|U|+\mathcal{R}^2N)$, which is closer to the basic PF algorithm [12].

V. Performance Analysis

We have analyzed the effectiveness of our proposed scheduler in 5G-LENA [8], a New Radio (NR) network simulator, designed as a pluggable module to ns-3. In the following, we have demonstrated the simulation scenario and model. Then, we determine the values of the configuration parameters $\mathcal{N}$ and $\vec{\delta}$. We have analyzed the performance analysis of our scheme based on the 3GPP recommended 5G RAN slicing scenarios [36].

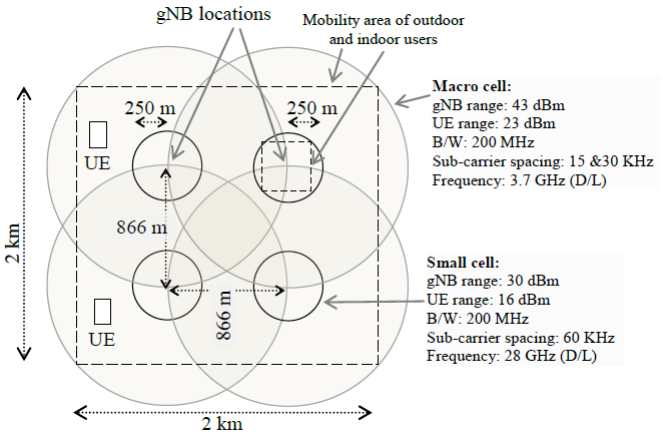

Fig. 2: Simulation scenario and typical cell-level parameters

TABLE III: 5G System Parameters

Parameter	*Value*
Duplex mode	TDD
Channel Quality Indicator (CQI) table	Table 5.2.2.1-3 of [31]
Propagation loss model	Outdoor urban and indoor urban propagation loss model of ns-3
MIMO mode	2×2
Modulation Coding Scheme (MCS) table	Table 5.1.3.1-1 and Table 5.1.3.1-2 of [31]
Mobility model	Random way-point model of ns-3

A. Simulation Scenarios

The simulation scenario and the cell-level typical parameters are shown in Fig. 2. The scenario is compatible with the 3GPP specified 5G Urban macro and Dense urban use cases [36]. Here, 80% users are at indoor coverage with a speed of 3km/h and the remaining 20% are at outdoor coverage moving with 100km/h speed. A UE can connect with more than one gNB while running multiple flows of different types, e.g., either MTC, or VoIP, or Video streaming. Hence multi-gNB scheduling becomes an effective way to increase the spectral efficiency of the system. This kind of cell-planning is generally seen in the Urban macro and Dense urban scenarios [36] to reduce the capital expenditures of the tenant operators. However, a scenario such as integrated terrestrial and non-terrestrial 5G and beyond networks [37] may not be appropriate for the study of our proposed scheme, since satellite links remain primarily unstable, making it impossible to calculate the data efficiency parameter (H_{ij}) correctly at each time interval. The QoS parameter values (such as delay) between cellular and satellite networks also differ significantly, making it impossible to accurately calculate the expiration time threshold (Δ). Other typical 5G system parameters considered are shown in Table III.

We have considered three slices: *Slice1* having MTC flows, *Slice2* having VoIP flows and *Slice3* having Video flows. For video flows, we have adopted H.265 video codec with frame rate of 30 fps and resolution of 3840×2160 4k UHD, requiring 15 Mbps bandwidth. The MTC flows run TCP echo application, implemented in ns-3. The service period for both VoIP and video flows are constant (120 s), whereas MTC

service period is continuous throughout the entire simulation time. The delay threshold for all kind of flows is set to 100 ms and the inter-packet delay threshold for VoIP and video flows are set to 40 ms while for MTC it is set to 1 ms. The PLR thresholds for MTC, VoIP and video flows are set to 10^{-2}, 10^{-2} and 10^{-6} respectively, where the throughput threshold are set to 20 Kbps, 12.65 Kbps and 15 Mbps respectively following constant packet generation at the source. The value of $d_{Server,UPF}$ is considered as exponentially distributed with mean 20 ms and $d_{UPF,UE}$ is fixed at 5 ms. To simulate the impact of delay variation of each packet due to diverse routing path, a delay variation of ± 5 ms is also considered. The traffic pattern of each MTC device can be defined by four states: OFF (no packet transmission), Periodic Update, Event-driven and Payload exchange [32]. The values of MTC traffic parameters, used to derive the mean packet arrival rate ($\overline{W}$), are $S = 4$, $\mathbf{P} = [(0, 0.5, 1, 1), (0.5, 0, 0, 0), (0.5, 0.5, 0, 0), (0, 0, 0, 0)]^T$, $\overline{L} = [1s, 0.5s, 0.5s, 0.5s]$ and $\overline{W} = [0, 55/s, 50/s, 0]$.

To measure the QoS performance of each slice, we have defined a new parameter, named service rate. Service rate of a flow is the ratio of the number of default averaging window (2s, as recommended by 3GPP) in which QoS requirement of the flow is met over the total number of default averaging window in the entire service duration of the flow. Service rate of a slice is the average of service rates of all the flows under that slice. The QoS performance of a flow is measured as the average goodput performance of the flow in the averaging window, where $goodput = throughput \times (1 - PLR)$. In the PLR calculation, we have considered both lost and expired packets in the network.

In order to eliminate the initial transient states, we have taken all simulation results after 1200s of simulation run. The FLS scheduling algorithm [9] assumes discrete-time filter coefficient, $C_k(m) \in [0..1]$ for k^{th} slice and m^{th} DTI [9], with $C_k(m+1) < C_k(m), m \in [1...10]$. One possible way to satisfy the condition is to set $C_k(1) = 1$ and $C_k(m+1) = C_k(m)/2$. The configuration parameters of iRSS and CBAP schemes are taken from in [10] and [11] respectively. The value of M is taken as the maximum between the end-to-end delay and inter-packet delay threshold in the corresponding simulation.

B. Determination of $\mathcal{N}$ and $\vec{\delta}$

In the objective function, given in Section III-D, we need to take samples of the random variable $\mathcal{Z}_i, \forall i$, which has one-to-one dependency with the random variable, ψ_i. We take samples of $\mathcal{Z}_i, \forall i$ by varying ψ_i. To simplify the sampling process, we have assumed same packet size for a particular flow/application and ψ_i follows an uniform distribution from 0 to one-forth size of the packet of that flow. Every sample has $|U|$ number of elements. Samples are generated in such a way that the inequality (27) is satisfied.

We have assumed the configuration parameters of (30) as $\xi = e^{-2}$, $\gamma = 2p$ which gives us $\mathcal{N} \geqslant p^{-2}$. So the number of samples actually defines the upper bound of the overflow probability.

In Fig. 3, we have shown the radio resource utilization in terms of spectral efficiency for different number of samples.

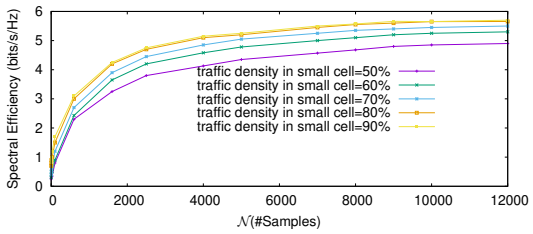

Fig. 3: Effect of number of samples on spectral efficiency

TABLE IV: Priority vector according to the target service rate

Cases different target service rate	**Scenario1** - priority vectors [1]	**Scenario2** - priority vectors [2]
Slice1 = 100%, *Slice2* = 100%, *Slice3* = $[90, 100]$%	$\delta_{Slice1} = 28$, $\delta_{Slice2} = 0.05$, $\delta_{Slice3} = 1$	$\delta_{Slice1} = 1.5$, $\delta_{Slice2} = 33$, $\delta_{Slice3} = 1$
Slice1 = 100%, *Slice2* = $[95, 100]$%, *Slice3* = $[90, 100]$%	$\delta_{Slice1} = 28$, $\delta_{Slice2} = 0.04$, $\delta_{Slice3} = 1$	$\delta_{Slice1} = 1.5$, $\delta_{Slice2} = 32.45$, $\delta_{Slice3} = 1$
Slice1 = $[99, 100]$%, *Slice2* = $[95, 100]$%, *Slice3* = $[90, 100]$%	$\delta_{Slice1} = 27.4$, $\delta_{Slice2} = 0.04$, $\delta_{Slice3} = 1$	$\delta_{Slice1} = 1.47$, $\delta_{Slice2} = 32.45$, $\delta_{Slice3} = 1$

[1] **Scenario1:** *Slice1*(MTC) > *Slice2*(VoIP) > *Slice3*(Video)
[2] **Scenario2:** *Slice1*(MTC) > *Slice2*(Video) > *Slice3*(VoIP)

If we increase $\mathcal{N}$, p will decrease, resulting in more accuracy in determining packet size in our objective function and more number of packets can be scheduled with the available resources. As a consequence, the spectral efficiency is increased while increasing $\mathcal{N}$. To increase the randomness of Channel Quality Indicator (CQI) reporting by UEs, we vary the traffic density in the small cell area. Instead of diverse values of $\mathcal{Z}_i$ in different cases, our algorithm accurately predicts the required amount of resources for every packets. Decreasing the traffic density decreases the number of CQI reporting as the chance of physical distribution in the larger rectangular area becomes higher and so the overall spectral efficiency also decreases as shown in Fig. 3. The spectral efficiency remains almost unchanged even we increase $\mathcal{N}$ beyond 10^4, irrespective of the traffic density. Therefore we can choose this value as a reference in our following simulation.

The lower bound of service rate of a slice can be fixed by tuning $\vec{\delta}$ for its packets. The values of $\vec{\delta}$ for different slices in case of different service rate requirements are given in Table IV. After setting δ to an appropriate value for all packets, our algorithm ensures that if we increase the load in the system, the service rate of a slice will not fall below its required level unless the service rates of all low priority slices become zero. In the following simulations, we set $\vec{\delta}$ as given in Table IV dynamically according to the slice priorities and their service rate requirements. If the system load increases beyond its capacity, it first affects the service rate of the lowest priority slice i.e., *Slice3*. We can prevent the service rate of *Slice3* from going down its required level by setting $\vec{\delta}$ of *Slice1* and *Slice2* as given in Table IV such that the service rates of *Slice1+2* still remain above the required level. This dynamic setting of $\vec{\delta}$ during scheduling gives us more flexibility to support heterogeneous service demands of multiple slices. In an online system, the priority vector $\vec{\delta}$ can also be dynamically

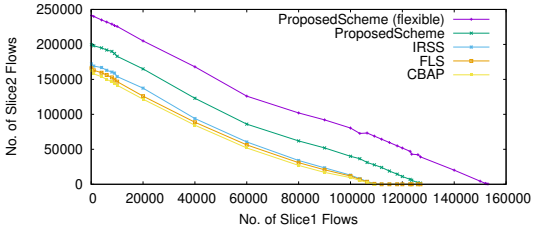

Fig. 4: Load distribution between Slice1 and Slice2

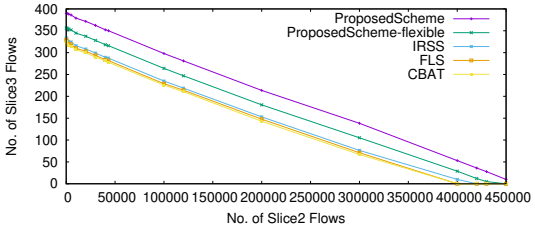

Fig. 5: Load distribution between Slice2 and Slice3

adjusted to match the required service rate as given in Table IV though online control systems.

C. Comparative Analysis

We have considered Scenario1 of Table IV and compared our proposed scheme with iRSS [10], CBAP [11] and FLS [9] schemes with respect to load of slices. In Fig. 4 and Fig. 5, we keep the load of *Slice3* and *Slice1* constant while varying the load of other two slices such that total radio resource of the system is fully utilized by all the slices. The iRSS scheme performs better than CBAP and FLS in terms of maximum load support. But, iRSS does not consider QoS of slices and static size of RBs while scheduling. It allocates a chunk of radio resources for an entire slice only, avoiding the mapping

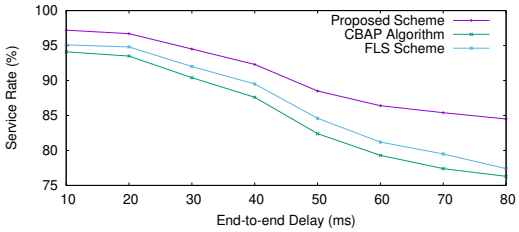

Fig. 6: Effect of end-to-end delay on slice performance

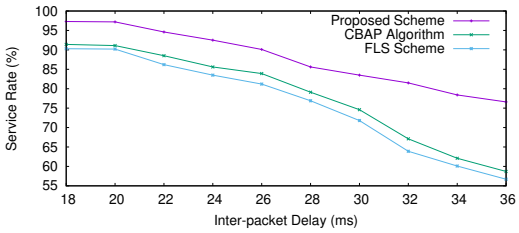

Fig. 7: Effect of inter-packet delay on slice performance

TABLE V: Values of QoS parameters

Slice no. (priority: top-down)	**Application supported**	**Parmaeters**
Slice1	URLLC	Throughput: 10 Mbps Delay: 1 ms Inter-packet delay: 1 ms PLR: 10^{-6} Service rate: 100%
Slice2	MTC	Throughput: 10 Mbps Delay: 100 ms Inter-packet delay: 1 ms PLR: 10^{-6} Service rate: 98%
Slice3	VoIP, Video	Throughput: 50 Mbps Delay: 100 ms Inter-packet delay: 40 ms PLR: 10^{-2} Service rate: 90%
Slice4	VoIP, Video	Throughput: 50 Mbps Delay: 100 ms Inter-packet delay: 40 ms PLR: 10^{-2} Service rate: 80%

of its selected radio resource into physical RB units, resulting into more resource demand than its predicted value. Whereas we have considered both of these issues in our system model and therefore our scheme performs much better than the iRSS scheme. CBAP and FLS are both QoS-aware scheme, but they are not designed to support optimum resource utilization in multi-gNB scenario. Also they consider only end-to-end delay as QoS parameter in their algorithm. Although FLS consider end-to-end delay, CBAP considers only framing delay of the packets and therefore, FLS performs slightly better than CBAP scheme.

A more realistic scenario in the perspective of 5G tenant-level operator is also considered. The required QoS of different slice flows are listed in Table V. Here a UE may subscribe to a particular slice to achieve the mentioned QoS requirements, independent of number of flows and types (VoIP/Video/etc.). The URLLC traffic is generated in a nearby cloud server and to maintain its end-to-end delay within 1 ms range we use the frame level delay reducing technique, CBAP [11]. In terms of supported load, our scheme outperforms other three schemes and the performance is almost same as shown in Fig. 4 and Fig. 5. As the Internet is highly dynamic, the delay and inter-packet delay also become highly unstable. Hence, measuring the performance of a slice in terms of service rate while varying the mean of exponentially distributed delays, would be very interesting. In Fig. 6 and Fig. 7, we have shown the comparative performance of three schemes namely, proposed scheme, CBAP and FLS. We do not consider iRSS as it is QoS-insensitive. As can be seen from Fig. 6 and Fig. 7, our proposed scheme outperforms CBAP and FLS. This is due to the fact that CBAP and FLS do not consider the inter-packet delay constraint. Besides FLS is a probabilistic approach and no further packet loss is guaranteed while scheduling current packets. The CBAP scheme only considers framing delay of packets without taking into account the end-to-end delay.

The above analysis shows that in a medium/high dense urban scenario, our scheme outperforms the existing scheduling

schemes of 5G RAN slicing. We are able to support spectral efficiency of approximately 5.6 bits/s/Hz, which is higher than the recommended value in the 5G 3GPP specification. The service rate of every slice can be controlled using $\vec{\delta}$ while the load of different slices vary dynamically with respect to time. The QoS of a higher priority slice can never go down a predefined service rate, even when the load of a lower priority slice is increased. Additionally, our scheme can handle packets having high delay and inter-packet delay where these two delays are highly dynamic due to diverse internet condition. In static condition with a low end-to-end delay and inter-packet delay, the proposed scheme performs almost same as others, but in the dynamic condition it performs much better than others due to their insensitivity to these delays.

VI. Conclusions

Our proposed scheme is ideal for the scenarios which need to ensure a minimum QoS for every slice, even though traffic arrival is random. In such scenarios, current literature adopts static resource reservation strategies; but these are costly approaches, as the overall spectral efficiency of the system is reduced. Our non-sequential allocation strategy gives every slice a fair chance to access a minimum amount of resources at each scheduling interval, even though the slices are unevenly loaded. The priority of a slice can also be changed during the scheduling process in order to provide more flexibility in service provisioning. For example, a video slice subscriber might need higher bandwidth while watching a live video than while downloading a video. Hence, the priority of the live video sub-slice can be upgraded, while the priority of the video sub-slice can be downgraded. Similarly, road safety messages may need to be prioritized over other vehicular applications within a Heterogeneous Vehicular Network (HetVNet) Slice from time to time. Thus flexible prioritization technique gives more options to the tenant-level operators to use fine-grained policies for their designated slices.

Currently, the new orientation of the flexible resource block configuration technique is becoming inefficient due to high power consumption, delay, and control overhead. In an overlay network environment, our proposed scheme can smartly map a packet to a resource block of appropriate size and increase resource utilization. In this context, in scenarios like massive and critical cellular IoT networks where most of the traffic flows contain packets of sizes much smaller than a radio resource unit, testing the performance of our proposed scheme can be an interesting future work.

On the other hand, while roadside units (RSUs) take charge of resource allocation in a HetVNet environment, dedicated resources are allocated for vehicle-to-vehicle (V2V) side links. It results in a poor data rate for V2V links. With the cooperation of RSUs, the cellular base station can apply our proposed scheme to allocate resources dynamically to V2V links according to their demands, thereby improving the data rate. To accommodate more radio resources for a single V2V link, reusing the same radio resources for different V2V links under the same cellular region can be a good option. Our proposed scheme does not take this possibility into account, and hence it will be an interesting future work if we also consider the reuse factor in our system model. During a handoff period, resource scheduling of vehicular traffic can be done by the nearby cellular base station instead of the short-range RSUs. Our proposed scheme can instantly schedule them by prioritizing them over other cellular traffic. Hence, a seamless handoff experience can be realized in a HetVNet environment. Therefore, the proposed resource scheduling scheme can be considered a significant contribution towards fulfilling 5G RAN slicing requirements.

However, some challenges may appear while deploying the proposed scheme in a real-world scenario. One of the key issues is the seamless mobility management requirement for UE between two SDN controllers while the allocation of radio resources is going on from both of them at different time intervals. This issue may become more complex if the RANs under the SDN controllers use different radio access technologies. Another issue to overcome is that the security and privacy policies of a particular slice must not be breached while allocating radio resources to another slice.

Appendix A

To prove that our problem is NP hard, we reduce the known NP hard subset sum problem to our problem. We use proof by contradiction. A special instance, Q of our problem can be obtained if we suppose that each RB has the same size independent of packet identity, i.e., $H_{ij} = H, \forall i, j$. The capacity constraint of Q can be expressed as $H \sum_{j=1}^{R} x_{ij} \geqslant b_i, \forall i$. Now, a decision version of Q can be stated as follows: Does there exist a solution of Q such that $\sum_{j=1}^{R} x_{ij} \geqslant b_i, \forall i$ and $\sum_{i=1}^{|U|} \delta_i y_i \geqslant D$, where D is an arbitrary profit value?

First, we can prove that Q is NP. The verification process is to compute $H \sum_{j=1}^{R} x_{ij} \geqslant b_i, \forall i$ and $\sum_{i=1}^{|U|} \delta_i y_i \geqslant D$ which takes polynomial time in proportion to the size of the input.

Second, we can show that there is a polynomial time reduction from the subset sum problem to Q. The subset sum problem can be given as $c_1, c_2, \ldots, c_{|U|}$ and V where we need to find $y_i, \forall i$ such that $\sum_{i=1}^{|U|} c_i y_i = V$. If there exists an efficient algorithm to solve Q, the total capacity constraint, i.e., $\sum_{i=1}^{|U|} b_i y_i \leqslant RH$, must be met. Now let us assume the following process of converting the subset sum problem to Q in polynomial time: $\delta_i = c_i, \forall i$, $b_i = c_i, \forall i$, $D = V$ and $H = V/R$. Now the solution of Q must satisfy these inequalities: $\sum_{i=1}^{|U|} c_i y_i \leqslant V$ and $\sum_{i=1}^{|U|} c_i y_i \geqslant V$ which implies that $\sum_{i=1}^{|U|} c_i y_i = V$. Therefore, $y_i, \forall i$ is the desired solution of the subset sum problem. This establishes the NP completeness of our problem.

Appendix B

Here, Our goal is to determine the value of $f(T, k)$. The distributions given in the Section III-C is used to compute $f(T, k)$.

Moreover the end-to-end delay and inter-packet delay vary depending upon routing path, hop distance, packet generation rate, queuing or processing delay, etc. In the 5G network, these delays are evenly distributed within a certain range and

therefore they can be modeled as a Uniform Distribution [38]. Following (2), it can be stated that Δ has a Uniform Distribution with a discrete probability distribution function:

$$\Pr\{\Delta = j\} = \begin{cases} \frac{1}{M}, & 0 \leqslant j \leqslant M \\ 0, & j < 0 \text{ or } j > M \end{cases} \tag{54}$$

Traffic arrival from a single voice application follows modified Poisson distribution [29]. As the sum of independent modified Poisson-distributed random variables also follows a modified Poisson distribution, the mean of the random variable for multiple voice applications, $\hat{Y}(= \sum_{i=1}^{n} y_i)$ is $n\mu$, where μ is the mean of the random variable, y_i. The value of μ can be obtained from (8).

At T^{th} DTI, a packet will be included in between $(T+1)$ to k^{th} subset if (Δ) is less than or equal to $(k-T)$. Using (54), the probability distribution of number of packets included up to k^{th} subset in T^{th} DTI can be derived as follows:

$$\Pr\{T,k\} = \sum_{y=x}^{\infty} \binom{y}{x} \big(\Pr\{\Delta \leqslant k-T\}\big)^x \big(\Pr\{\Delta > k-T\}\big)^{y-x} \tag{55}$$

$$\Pr\{\hat{Y} = y\} = e^{-\gamma}\frac{\gamma^x}{x!} \tag{56}$$

where $\gamma = \frac{n\mu(k-T)}{M}$. It is also following the modified Poisson distribution. So, the event of packet arrival from T to $(k-1)^{\text{th}}$ DTI would follow the summation of the above modified Poisson distribution. Therefore, $f(T,k)$ can be calculated as follows,

$$f(T,k) = \sum_{i=T}^{k-1} \frac{n\mu(k-i)}{M} = \frac{n\mu(k-T)(k-T+1)}{2M} \tag{57}$$

The inter-arrival time of video packets follows Pareto distribution [39]. As Pareto process does not hold summation and multiplication property like modified Poisson, it would be very difficult to determine the joint probability distributions. So, first we calculate the mean number of packets that will be included up to k^{th} subset at T^{th} DTI as $\frac{k-T}{M}\frac{1}{\mu}$. Following up to $(k-1)^{\text{th}}$ DTI, $f(T,k)$ can be calculated as follows,

$$f(T,k) = \frac{(k-T)(k-T+1)}{2\mu M} \tag{58}$$

where the Pareto parameters z_m and α can be estimated directly from the data samples (see [40] for example).

Similarly, $f(T,k)$ can be calculated for the SMM as:

$$f(T,k) = \frac{(k-T)(k-T+1)n\overline{W}}{2M} \tag{59}$$

So, the mean number of packets, which will arrive up to k^{th} subset in between T and $(k-1)^{\text{th}}$ DTI can be calculated using (57), (58) and (59) for voice, video and MTC applications respectively.

APPENDIX C

Proposition 3. *If mean number of packets arrive in each DTI and mean number of packets included in each subset, then the transmission need $(d_j^{(i)})$, calculated using (51), becomes a decreasing function with respect to j.*

Proof. Let us assume that k^{th} and s^{th} indexed subsets are the highest selection metrics according to (51) at T^{th} and $(T+1)^{\text{th}}$ DTI respectively, where $k > T$ and $s > T+1$. We can also assume the mean number of packets included in each subset at each DTI as m^i. According to (52), $m^{(voice)} = n\mu$ and $m^{(video)} = \frac{1}{\mu}$. So, we can write the following equations.

$$d_T^{(i)} = \varphi_{T,k}^{(i)} + \frac{(k-T+1)m^{(i)}}{2M} \tag{60}$$

$$d_{T+1}^{(i)} = \varphi_{T+1,s}^{(i)} i + \frac{(s-T)m^{(i)}}{2M} \tag{61}$$

As k^{th} indexed subset is the highest metric at T^{th} DTI, it will be greater than or equal to the selection metric calculated for s^{th} subset at that DTI. The inequality can be written as,

$$d_T^{(i)} \geqslant \varphi_{T,s}^{(i)} + \frac{(s-T+1)m^{(i)}}{2M} \tag{62}$$

The value of $\varphi_{T+1,s}^{(i)}$ can be calculated as follows:

$$\varphi_{T+1,s}^{(i)} = \frac{(s-T)\varphi_{T,s}^{(i)} + (s-T-1)\frac{m^{(i)}}{M} - d_T^{(i)}}{s-T-1} \tag{63}$$

By putting the value of $\varphi_{T+1,s}^{(i)}$ in (61), we can calculate the value of $d_{T+1}^{(i)}$. Applying the inequality (62) and $s > (T+1)$, the following inequality is proved.

$$d_T^{(i)} > d_{T+1}^{(i)} \tag{64}$$

Similarly, we can also prove, $d_{T+1}^{(i)} > d_{T+2}^{(i)}$, $d_{T+2}^{(i)} > d_{T+3}^{(i)}$ and so on. It concludes that the transmission need $d_j^{(i)}$, is a decreasing function with respect to j under the above-mentioned assumption. □

REFERENCES

[1] A. R. Hossain and N. Ansari, "Priority-Based Downlink Wireless Resource Provisioning for Radio Access Network Slicing," *IEEE Transactions on Vehicular Technology*, vol. 70, no. 9, pp. 9273–9281, 2021.

[2] 5GPP, "View on 5G Architecture," 5GPPP Architecture Working Group, Tech. Rep., October 2021. [Online]. Available: https://5g-ppp.eu/wp-content/uploads/2021/11/Architecture-WP-V4.0-final.pdf

[3] W. K. Seah, C.-H. Lee, Y.-D. Lin, and Y.-C. Lai, "Combined Communication and Computing Resource Scheduling in Sliced 5G Multi-Access Edge Computing Systems," *IEEE Transactions on Vehicular Technology*, vol. 71, no. 3, pp. 3144–3154, 2022.

[4] B. Picano, R. Fantacci, and Z. Han, "Nonlinear Dynamic Chaos Theory Framework for Passenger Demand Forecasting in Smart City," *IEEE Transactions on Vehicular Technology*, vol. 68, no. 9, pp. 8533–8545, 2019.

[5] R. Fantacci and B. Picano, "End-to-End Delay Bound for Wireless UVR Services Over 6G Terahertz Communications," *IEEE Internet of Things Journal*, vol. 8, no. 23, pp. 17 090–17 099, 2021.

[6] B. Pagnoncelli, S. Ahmed, and A. Shapiro, "Sample Average Approximation Method for Chance Constrained Programming: Theory and Applications," *Journal of Optimization Theory and Applications*, vol. 142, pp. 399–416, 08 2009.

[7] R. Guerin, H. Ahmadi, and M. Naghshineh, "Equivalent capacity and its application to bandwidth allocation in high-speed networks," *IEEE Journal on Selected Areas in Communications*, vol. 9, no. 7, pp. 968–981, 1991.

[8] N. Patriciello, S. Lagen, B. Bojovic, and L. Giupponi, "An E2E simulator for 5G NR networks," *Simulation Modelling Practice and Theory*, vol. 96, p. 101933, 2019. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1569190X19300589

[9] G. Piro, L. A. Grieco, G. Boggia, R. Fortuna, and P. Camarda, "Two-Level Downlink Scheduling for Real-Time Multimedia Services in LTE networks," *IEEE Transactions on Multimedia*, vol. 13, no. 5, pp. 1052–1065, Oct. 2011.

[10] M. Yan, G. Feng, J. Zhou, Y. Sun, and Y.-C. Liang, "Intelligent Resource Scheduling for 5G Radio Access Network Slicing," *IEEE Transactions on Vehicular Technology*, vol. 68, no. 8, pp. 7691–7703, 2019.

[11] Y. Boujelben, "Scalable and QoS-Aware Resource Allocation to Heterogeneous Traffic Flows in 5G," *IEEE Internet of Things Journal*, vol. 8, no. 20, pp. 15 568–15 581, 2021.

[12] F. Capozzi, G. Piro, L. A. Grieco, G. Boggia, and P. Camarda, "Downlink Packet Scheduling in LTE cellular networks: Key design issues and a survey," *IEEE Communications Surveys Tutorials*, vol. 15, no. 2, pp. 678–700, Sep. 2013.

[13] G. Monghal, K. I. Pedersen, I. Z. Kovacs, and P. E. Mogensen, "QoS oriented time and frequency domain packet schedulers for the utran long term evolution," in *Proc. VTC Spring 2008 - IEEE Vehicular Technology Conf*, May 2008, pp. 2532–2536.

[14] Y. C. Wang and T. Y. Tsai, "A Pricing-Aware Resource Scheduling Framework for LTE networks," *IEEE/ACM Transactions on Networking*, vol. 25, no. 3, pp. 1445–1458, Jun. 2017.

[15] A. Ksentini and N. Nikaein, "Toward Enforcing Network Slicing on RAN: Flexibility and Resources Abstraction," *IEEE Communications Magazine*, vol. 55, no. 6, pp. 102–108, June 2017.

[16] T. Guo and A. Suárez, "Enabling 5G RAN Slicing with EDF Slice Scheduling," *IEEE Transactions on Vehicular Technology*, pp. 1–1, 2019.

[17] S. Chatterjee, M. J. Abdel-Rahman, and A. B. MacKenzie, "On Optimal Orchestration of Virtualized Cellular Networks With Downlink Rate Coverage Probability Constraints," *IEEE Transactions on Wireless Communications*, vol. 19, no. 7, pp. 4378–4393, 2020.

[18] S. Chatterjee, M. J. Abdel-Rahman, and A. B. MacKenzie, "On Optimal Orchestration of Virtualized Cellular Networks With Statistical Multiplexing," *IEEE Transactions on Wireless Communications*, vol. 21, no. 1, pp. 310–325, 2022.

[19] A. Sharifian, R. Schoenen, and H. Yanikomeroglu, "Joint Realtime and Nonrealtime Flows Packet Scheduling and Resource Block Allocation in Wireless OFDMA networks," *IEEE Transactions on Vehicular Technology*, vol. 65, no. 4, pp. 2589–2607, Apr. 2016.

[20] W. K. Lai and C.-L. Tang, "QoS-aware downlink packet scheduling for lte networks," *Comput. Netw.*, no. 7, pp. 1689–1698, May 2010.

[21] K. Kaewmongkol, A. Jansang, and A. Phonphoem, "Delay-aware with resource block management scheduling algorithm in LTE," in *Proc. Int. Computer Science and Engineering Conf. (ICSEC)*, Nov. 2015, pp. 1–6.

[22] P. Rahimi, C. Chrysostomou, H. Pervaiz, V. Vassiliou, and Q. Ni, "Joint Radio Resource Allocation and Beamforming Optimization for Industrial Internet of Things in Software-Defined Networking-Based Virtual Fog-Radio Access Network 5G-and-Beyond Wireless Environments," *IEEE Transactions on Industrial Informatics*, vol. 18, no. 6, pp. 4198–4209, 2022.

[23] Q. Zheng, K. Zheng, H. Zhang, and V. C. M. Leung, "Delay-Optimal Virtualized Radio Resource Scheduling in Software-Defined Vehicular Networks via Stochastic Learning," *IEEE Transactions on Vehicular Technology*, vol. 65, no. 10, pp. 7857–7867, 2016.

[24] M. Hu, Y. Chang, Y. Sun, and H. Li, "Dynamic slicing and scheduling for wireless network virtualization in downlink LTE system," in *2016 19th International Symposium on Wireless Personal Multimedia Communications (WPMC)*, Nov 2016, pp. 153–158.

[25] Q. Ye, J. Li, K. Qu, W. Zhuang, X. S. Shen, and X. Li, "End-to-End Quality of Service in 5G Networks: Examining the Effectiveness of a Network Slicing Framework," *IEEE Vehicular Technology Magazine*, vol. 13, no. 2, pp. 65–74, June 2018.

[26] Y. Prathyusha and T.-L. Sheu, "Coordinated resource allocations for eMBB and URLLC in 5G communication networks," *IEEE Transactions on Vehicular Technology*, vol. 71, no. 8, pp. 8717–8728, 2022.

[27] M. Laner, P. Svoboda, P. Romirer-Maierhofer, N. Nikaein, F. Ricciato, and M. Rupp, "A comparison between one-way delays in operating HSPA and LTE networks," in *Proc. Ad Hoc and Wireless Networks (WiOpt) 2012 10th Int. Symp. Modeling and Optimization in Mobile*, May 2012, pp. 286–292.

[28] N. Elkins, M. Ackermann, A. Deshpande, T. Pecorella, and A. Rashid, "IPv6 Performance and Diagnostic Metrics Version 2 (PDMv2) Destination Option," Internet Engineering Task Force, Internet-Draft draft-elkins-ippm-encrypted-pdmv2-02, Feb. 2022, work in Progress. [Online]. Available: https://datatracker.ietf.org/doc/draft-elkins-ippm-encrypted-pdmv2/02/

[29] S. Zhang, Z. Zhao, H. Guan, and H. Yang, "A modified poisson distribution for smartphone background traffic in cellular networks," *International Journal of Communication Systems*, vol. 30, no. 6, 2016.

[30] I. Al Ajarmeh, J. Yu, and M. Amezziane, "Framework of Applying a Non-homogeneous Poisson Process to Model VoIP Traffic on Tandem Networks," in *Proc. 10th WSEAS International Conference on Applied Informatics and Communications*. World Scientific and Engineering Academy and Society (WSEAS), 2010, pp. 164–169.

[31] 3GPP, "TSG RAN; NR; Physical layer procedures for data," 3rd Generation Partnership Project (3GPP), Technical Specification (TS) 38.214, December 2019, version 16.0.0. [Online]. Available: https://portal.3gpp.org/desktopmodules/Specifications/SpecificationDetails.aspx?specificationId=3216

[32] N. Nikaein, M. Laner, K. Zhou, P. Svoboda, D. Drajic, M. Popovic, and S. Krco, "Simple Traffic Modeling Framework for Machine Type Communication," in *ISWCS 2013; The Tenth International Symposium on Wireless Communication Systems*, Aug 2013, pp. 1–5.

[33] S. M. Hasan, M. A. Hayat, and M. F. Hossain, "On the downlink SINR and outage probability of stochastic geometry based LTE cellular networks with multi-class services," in *2015 18th International Conference on Computer and Information Technology (ICCIT)*, 2015, pp. 65–69.

[34] J. Luedtke and S. Ahmed, "A Sample Approximation Approach for Optimization with Probabilistic Constraints," *SIAM Journal on Optimization*, vol. 19, no. 2, pp. 674–699, 2008.

[35] F. Furini, M. Monaci, and E. Traversi, "Exact approaches for the knapsack problem with setups," *Computers & Operations Research*, vol. 90, pp. 208–220, 2018.

[36] J. Navarro-Ortiz, P. Romero-Diaz, S. Sendra, P. Ameigeiras, J. J. Ramos-Munoz, and J. M. Lopez-Soler, "A Survey on 5G Usage Scenarios and Traffic Models," *IEEE Communications Surveys & Tutorials*, vol. 22, no. 2, pp. 905–929, 2020.

[37] F. Rinaldi, S. Pizzi, A. Molinaro, A. Iera, and G. Araniti, "Cooperative resource allocation in integrated terrestrial/non-terrestrial 5G and beyond networks," in *GLOBECOM 2020 - 2020 IEEE Global Communications Conference*, 2020, pp. 1–6.

[38] B. Coll-Perales, M. Lucas-Estañ, T. Shimizu, J. Gozalvez, T. Higuchi, S. Avedisov, O. Altintas, and M. Sepulcre, "End-to-End V2X Latency Modeling and Analysis in 5G Networks," *IEEE Transactions on Vehicular Technology*, pp. 1–15, 2022.

[39] X. Yang, "Designing traffic profiles for bursty Internet traffic," in *Proc. IEEE Global Telecommunications Conf. GLOBECOM '02*, vol. 3, Nov. 2002, pp. 2149–2154 vol.3.

[40] M. Rytgaard, "Estimation in the Pareto distribution," *ASTIN Bulletin*, vol. 20, no. 2, pp. 201–216, 1990.

Manoj Kumar Rana a Research Assistant Professor in the department of Computing Technologies under the College of Engineering and Technology of the SRM Institute of Science and Technology, Kattankulathur, Chennai, received his B.Tech. degree in Computer Science and Engineering from Kalyani Govt. Engineering College, West Bengal, India, and his M.Tech. and Ph.D. degrees from Jadavpur University, both in the School of Mobile Computing and Communication. He has published and also worked as a reviewer for several esteemed international journals and conferences. His research interests focus on 5G and beyond network communication systems, network simulation, and the application of machine learning to networking systems, etc. Dr. Rana received the prestigious scholarship from TEQIP, phase II, under the MHRD of the Govt. of India, during his Ph.D. and also received the Google Summer of Code (GSoC) scholarship for his significant contribution to the Network Simulator 3 (NS-3) project in 2017. He is a regular member of the NS-3 organization and has been mentoring for NS-3 in GSoC since 2021.

This article has been accepted for publication in IEEE Transactions on Vehicular Technology. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TVT.2023.3327874

Tommaso Pecorella (Senior Member, IEEE) received the Ph.D. and M.Sc. degrees in electronic engineering (telecommunications track) from the Department of Information Engineering, University of Florence, Italy, in 2000 and 1996, respectively. From 2001 to 2007, he was a Researcher at Consorzio Nazionale Interuniversitario per le Telecomunicazioni (CNIT). Since November 2007, he has been a Tenure Track Assistant Professor with the Department of Information Engineering, University of Florence. In 2018 and 2019, he was also a Visiting Professor with the University of Saint Louis, Missouri, USA. He is the author of more than 90 publications between conference papers and journals. His research interests focus on IoT communication systems, network security, and application of machine learning to networking systems. Dr. Pecorella received the Best Paper Award at the IEEE GLOBECOM 2016, and in 2021 got the Italian Habilitation (Abilitazione Scientifica Nazionale) for Associate Professorship in Telecommunication Engineering.

Bhaskar Sardar a professor with the department of Information Techonology, Jadavpur University, Kolkata, India. He received his BE, ME, and PhD degrees in computer science and engineering from Jadavpur University, Kolkata, India. He has more than 20 years of teaching and research experience. He has published numerous research articles in reputed conferences and journals. His current research interests include TCP for wired/wireless networks, distributed mobility management techniques, routing protocols, resource optimization in 5G and beyond.

Rama Rao Thipparaju a Professor at SRM Institute of Science and Technology, Kattankulathur, Chennai India and has long-standing research experience in Radio Communications and Wireless Networks. Earlier, he worked at Aalborg University, Denmark; at Universidad Carlos III de Madrid, Spain, and at the University of Sydney, Australia, as a Visiting Professor. Prof. Rama Rao received significant funding from the DST / DRDO / ISRO, Government of India to develop Millimeter-wave (MmW) communications and Inter-satellite radio link investigations at MmW; Developed wearable antennas for device applications and involved in developing Photonic Wireless Communications for 5G beyond communications. He is a Sr. Member of IEEE, and a Fellow of IETE, India.

Debashis Saha an Endowed Chair Professor in the MIS area in IIM Calcutta, received his B.E. (Hons) degree from Jadavpur University, Kolkata, India, and M.Tech. and Ph. D. degrees from the Indian Institute of Technology (IIT), Kharagpur, all in Electronics and Communication Engineering. He has supervised/co-supervised 21 doctoral theses, published about 350 research papers in various international journals and conferences, and directed 4 funded projects on Telecom/IT. He has co-edited three books, and coauthored several book chapters, a monograph and five books He was the Co-Editor-in-Chief of the International Journal of Business Data Communications and Networking (IJBDCN) [2009-2013]. He had served on the editorial board of select international journals, and is a regular reviewer of several top-rated journals. Prof. Saha has won several best paper awards in various conferences, was the recipient of the prestigious Career Award for Young Teachers from AICTE, Government of India, and is a Fellow of West Bengal Academy of Science and Technology (WAST), Senior Life Member of Computer Society of India, and Senior Member of IEEE.