Predictive Dynamic Scaling Multi-Slice-in-Slice-Connected Users for 5G System Resource Scheduling

Sharvari Ravindran, Saptarshi Chaudhuri, Jyotsna Bapat, and Debabrata Das

Networking and Communication Research Lab
International Institute of Information Technology Bangalore (IIITB), India

Email: Sharvari.R@iiitb.org, saptarshi.chauduri@iiitb.org, jbapat@iiitb.ac.in, ddas@iiitb.ac.in

Abstract—Network slicing is an effective 5G concept for improved resource utilization and service scalability tailored to users (UEs) requirements. According to the standardization, 5G system should support UEs through specification of its heterogenous requirements such as, high data rates, traffic density, latency, reliability, UE density, system efficiency and service availability. These requirements are specified as Service Level Agreements (SLAs) between UEs and network operators resulting in increased interest to develop novel challenging mechanisms for improved interference-free resource efficient performances. An emerging concept that enables such demanding SLA assurances is Open-Radio Access Network (O-RAN). In this paper, we study a novel resource scheduling problem for UE services conditioned on other services in network slicing. To improve system performance, we propose that UEs are connected across network slices to serve several applications. UEs of similar service classes (defined by SLAs) are grouped to form optimized slice-in-slice category(s) within network slice(s). We propose novel Predictive Dynamic Scaling Multi-UE service specific System Resource Optimized Scheduling (DMUSO) algorithm(s). Multi-objective multi-constraint optimization problems (MOP) are formed to learn the dynamic system resource allocation and throughput for UE services conditioned on new services entering the network slice. An epsilon-constraint line search algorithm is presented to estimate UE service bandwidth. Using the theoretical models, DMUSO forms optimal slice-in-slice categories and estimates the maximum dynamic additional slice-in-slice categories, throughput served across network slices. Finally, compared to state-of-the-art literatures, DMUSO guarantees UEs SLAs with 4.4 and 7.5 times performance gains.

Index Terms—slice-in-slice category, system resources, scheduling.

I. INTRODUCTION

In 5G, the concept of network slicing has been introduced for better resource utilization efficiency, flexibility and support of fast growing over the top (OTT) services. Network slice (or slice) is an independent dynamically created logical entity with a set of functions, services and system resources. Each network slice allows operators to provide services tailored to user equipments (UEs) requirements over a single Radio Access Network (RAN). According to the 3GPP standardization, 5G and Beyond should support UEs multiple needs such as, data rates, service availabilities, latency, reliability and traffic densities with resources availability. These UE specific capabilities form the basis for developing flexible and intelligent mechanisms to support heterogeneous services, multi-connectivity, on-demand service deployment and ensure guaranteed Service Level Agreement (SLAs). The advent of the Open-Radio Access Network (O-RAN) base station architecture with support of intelligent models achieves efficient system optimization. A UE entering the network gets associated with the slices in RAN and its services. To ensure guaranteed SLAs, the RAN slice scenario starts with UE’s requirements [1]. The RAN is accordingly controlled by monitoring UEs long-term trends and patterns to achieve system resource allocation. However, owing to the surge in UE capacity, varying service data rates, fluctuating traffic and real-time channel variation, the dynamic network slice management is critical and challenging. In view of this, new mechanisms need to be developed which requires allocating resources optimally to accommodate future time-varying service demands [3].

The relevant recent literatures of medium access control (MAC) scheduling and slicing in 5G have been reviewed. In [2], the authors have developed a resource scheduling mechanism to minimize power consumption in O-RAN slicing. The scheduling is achieved by mapping the slices to services and resources to slices, as a mixed integer optimization problem. The authors in [6] have projected a RAN slicing method that flexibly allocates resources using deep reinforcement learning. The method ensures optimal allocation independent of the number of slices, without overprovisioning resources. The work in [15][17] have focused on priority-based resource allocation in network slices (RANS algorithm [15]) and resource optimized scheduling strategy (ROSS algorithm [16]) for evaluating users priority depending on the application. The work in [4][13] have achieved resource allocation considering the SLA contracts, small-time scale network dynamics, network states and traffic data across slices. In [14], authors have formulated a fairness-based multi-resource framework based on Ordered Weighted Average (MOWA algorithm) to satisfy users’ requirements, where the relation between resources is linear. However, the work does not consider SLA constraints to achieve resource scheduling.

Most of the prior work have relied on pre-negotiated SLAs, which monitor the responsibilities of UEs and define the fractional resource allocation. A reasonable approach to satisfy SLAs is triggering a reconfiguration process to allocate additional resources to UEs due to the uncertain demands.
However, above existing methods of system resource scheduling will fail to handle,

1) Optimal resource allocation for UEs to prevent over-provisioning the system due to UEs fluctuating demands.

2) Network scalability support for UEs and services. The single hardware of a deployed O-RAN base station (gNodeB) of network slice must be scalable and intelligent to accommodate UEs services owing to the heterogeneous complexity of wireless systems.

3) Analysing the channel interference across UEs with respect to optimal UE-specific system resource allocation.

4) With the concept of dynamic infrastructure sharing [8], heterogeneous complexity of wireless systems.

Considering all the aforementioned challenges, ensuring quality by learning the service demands, data rates, resource, bandwidth, efficiency, transmission rates and channel related SLAs have to be yet addressed extensively for a UE-specific level in network slicing. In view of this, to improve system performance, we propose the concept of multi-slice-in-slice-connected UEs to serve applications. Fig. 1 depicts a novel representation of multi-slice-in-slice-connected UEs across network slices 1,...,N. Single and Multiple UEs gets associated with network slices and serves several different applications. The UE(s) services having similar SLAs ranges (defined by service classes) are grouped to form multiple slice-in-slice categories within each network slice. These slice-in-slice categories operate within the resource envelope of its encompassing slice. We summarize our novel contributions,

1) A predictive dynamic scaling multi-slice-in-slice-connected UE(s) services for System resource Optimized Scheduling (DMUSO) algorithm across time intervals.

2) An algorithm for formation of optimized slice-in-slice service categories across network slice(s).

3) An algorithm to dynamically learn interference-aware UE service system resources and throughput conditioned on new services. The predicted service throughput estimates the maximum additional slice-in-slice categories and throughput for network slice(s).

4) An algorithm of learning the Pareto optimal bandwidth of UE(s) service(s) for resource scheduling.

5) Extensive system-level simulations using our novel analytical models to validate DMUSO performance. We show that the resources are allocated optimally leading to improved system performance gains compared to state-of-the-art methodologies.

The paper is structured as follows: Section II highlights the multi-slice-in-slice system parameters. Section III describes our proposed DMUSO algorithm and extensive analytical models. Section IV provides the simulation parameters. Section V describes the results and discussion. Finally, Section VI highlights the conclusions and future scope.

II. MULTI-SLICE-IN-SLICE SYSTEM PARAMETERS

This section describes the novel concepts of multi-slice-in-slice categories and the key implementation phases to achieve optimal and efficient system resource scheduling.

Definition 1: A $ith$ multi-slice-in-slice category is an amalgamation of UE(s) similar SLAs, i.e., service class defined by

$$\bigcup_{i=1}^{E(n)} T^{(n)} + x, \bigcup_{n=1}^{E(n)} E^{(n)} + y$$

where $T^{(n)}$, $E^{(n)}$ are the throughput, spectral efficiency (section III) and $\delta_{E^{(n)}}$, $\delta_{E^{(n)}}$ are their respective variances. In the O-RAN architecture [10], our work of scheduling focuses on interaction between MAC Open distributed unit (O-DU)-Low-PHY [7] Open Radio unit (O-RU) for UE specific channel interference aware resource allocation. The key implementation phases are,

1) Model capability query comprising of information for model training, such as system resources, network slicing infrastructure for implementing models and QoS-SLA constraint (Section III). In the RAN, the system resources $\forall$ network slice(s) $N$ is defined as; radio ($R^{(N)}$), virtualized ($V^{(N)}$) and transport ($T^{(N)}$) (shown in Fig. 1), in vector form,

$$\{R^{(N)}, V^{(N)}, T^{(N)}\} = \{|R_1^{(N)}, ..., R_S^{(N)}|, |V_1^{(N)}, ..., V_S^{(N)}|, |T_1^{(N)}, ..., T_S^{(N)}|\}$$

where $R_S^{(N)}, V_S^{(N)}, T_S^{(N)}$ are the resources for $S$th slice-in-slice category, decomposed into $\{R_1^{(N)}, ..., R_U^{(N)}\}, |V_1^{(N)}, ..., V_U^{(N)}|, |T_1^{(N)}, ..., T_U^{(N)}|$ for UE(s) service(s) $1, ..., U(S)$. In this work, we achieve Scheduling with respect to Radio resources (S-RS) and Transport resources (S-TS).

2) Model selection and training at every transmission time interval (TTI) of 1 msec. To achieve scheduling, we propose,

$$\{R^{(N)}, V^{(N)}, T^{(N)}\} = \{U_{(S)}, S_{(S)}, \text{constant}, E^{(N)}\}$$

system resources is learnt as a function of multiple network parameters $\rightarrow$ bandwidth ($B$), user throughput ($T_U$), spectral efficiency ($S_e$), interference ($I_e$) and channel variation ($C_{Q_i}$).

3) Model deployment and inference. Depending on the SLAs, scheduler maps a dynamic fraction ($\beta(.)$) of system resources for UE service(s). The resource allocation is formulated as a 2-tuple ($D_{n+k}^{(N)} = \sum_{i=1}^{m+k} \beta(R_{i,m+k}, T_{i,S})$, $\{R^{(N)}, T^{(N)}\}$ = $\sum_{m+k=1}^{S} D_{n+k}^{(N)}$) where $D_{n+k}^{(N)}$ is the demand of an $(m+k)^{th}$ slice-in-slice category (Defn. 3). Inference counters are further
used for tracking UE services \((U(m+k))\) at a slice-in-slice-level and maximum slice-in-slice categories \((S)\) served.

4) Model performance monitoring and updating comprising of aggregation of system metrics across UE services. Depending on environment feedback (channel conditions, new UEs services entering the slice), the performance evaluation module informs the system to update the current model. Subsequently, model is re-trained and deployed for inference.

III. PROPOSED PREDICTION-BASED DMUSO ALGORITHM AND ANALYTICAL MODEL

This section is divided into three sub-sections. First sub-section focuses on the system model for UEs within network slice(s). Second sub-section focuses on dynamically learning the system resources and throughput of UE service(s); Third sub-section focuses on estimating the Pareto optimal bandwidth of UE service(s) and complete DMUSO algorithm.

A. System model for multi-slice-in-slice-connected UE(s)

We consider the RAN slices \(N_1, \ldots, N_Q\) with \(V_1^{(N_1)}, \ldots, V_U^{(N_1)}\) being the multi-slice-in-slice-connected UE services (or referred to as UE services) per \(N_1\). If \(U\) be the cardinality of set of UE services within \(N_1\), \(U = |V|\).

**Definition 3:** Let \(m\) and \(k\) denote the existing and new slice-in-slice categories (formed) within a network slice. Let \(i_{m+k} = 1, \ldots, U(m+k)\) (where \(U(m+k)\)): Total number of services) denote the service(s) \(i\) in \((m+k)^{th}\) slice-in-slice category, whose capacity is \(l_{m+k}\). Then, capacity of network slice is \(\hat{l} = \sum_{m+k=1}^{S} l_{m+k}(i)\). Consider an orthogonal frequency domain multiplexing-based multiple input multiple output downlink transmission system (gNodeB to UE). Let \(u\) denote the index of a UE. The received signal [12] for \(u^{th}\) UE from gNodeB of a network slice \(N\) is:

\[
y^{(N)}_u = \sqrt{t_u} h_u^H e_u x_u + \sum_{q(i) \in N} \sqrt{t_q} h_q^H e_q x_q + n_u\tag{2}
\]

where \(t_u, t_q\) are \(w^{th}, q^{th} (q \neq i)\) UEs transmit powers allocated by O-RUs [2], \(h_u, h_q\) are channel vectors between UE \(u\), UEs \(q\) and O-RUs, \(e_u, e_q\) are unit norm beamforming [8] vectors from O-RUs to UEs, \(x_u, x_q\) are information symbols for UE \(u\), UEs \(q\), \(n_u \sim CN(0, N_0 b_{i,m+k})\) is the receiver noise. Between UE \(u\) and gNodeB, \((2) \Rightarrow \gamma_{i,m+k} = \frac{|h^H_u e_u|^2 t_u}{N_0 b_{i,m+k} + \sum_{q(i) \in N} |h^H_q e_q|^2 t_q}\tag{3}\)

is the received signal to interference noise ratio for \(u\), where \(b_{i,m+k}\) is the estimated bandwidth (S-TS) of each \(i \in u\) in \(N\).

B. Learning-based dynamic method of prediction of S-RSs and service throughput for UE(s) services

While achieving system resource scheduling, it is necessary to know how slice-in-slice categories are formed. The assumption is we deliberately create room for formation of a new \((k)\) slice-in-slice category within a network slice. Let \(u_{i,m+k}(t)\) denote the UEs \(i^{th}\) service throughput [3],

\[
u_{i,m+k}(t) = \frac{f_{d} n_{i,m+k}}{\Delta t}\tag{4}\]

where \(f_{d} = 2^{\mu} \times 15 kHz \times 12\), \(\mu\) is numerology [5], \(n_{i,m+k}\) is the spectral efficiency and \(\Delta t\) is TTI (1 msec). In this work, we consider the channel of UE(s) to be Rayleigh. Hence, using the concept of exponential utility function [3], \(\eta_{i,m+k}\) is modeled as \(e^{\beta(r_{i,m+k}\kappa_{i,m+k})}\), \(\beta\) is a constant, \(r_{i,m+k}\) is the S-RS allocation for UE service and \(s_{i,m+k}\) is the average SNR. We define the objective \((O)\) as maximization of throughput across UE services, such that sum of \(r_{i,m+k}\) across UE services should be less than the maximum S-RSs per network slice \(r_{\text{max}}\).

\[
\begin{align*}
O : & \max \sum_{m+k=1}^{S} \sum_{i=1}^{U(m+k)} u_{i,m+k}(t) \\
\text{Subject to,} & : \sum_{m+k=1}^{S} r_{i,m+k} \leq r_{\text{max}}
\end{align*}
\]

(5)

Generally, as SNR increases, throughput increases. Due to the Rayleigh channel of UEs, \(s_{i,m+k} \sim \frac{1}{\sigma^2_v} e^{-\frac{u}{\sigma^2_v}}\) of mean \(\sigma^2_v\).

Algorithm 1 Learnable S-RS and UE service throughput

1: \(i_{m+k} = i_{\text{current}}(1) \rightarrow i^{th}\) service \((m+k)\)
2: for \(k \neq 0\) do
3: \(\{\lambda_{1, ST}\} \leftarrow \{\lambda_{1, ST}\}_{i=U(m+k)}\)
4: Go to Step 6
5: end for
6: for \(i_{m+k}\) do
7: \(\{\lambda_{1, ST}\} \leftarrow \{\lambda_{1, ST}\}_{i=U(m+k)}\)
8: for \(i_{m+k}\) do
9: \(\{\lambda_{1, ST}\} \leftarrow \{\lambda_{1, ST}\}_{i=U(m+k)}\)
10: end for
11: \(\{\lambda_{1, ST}\} \leftarrow \{\lambda_{1, ST}\}_{i=U(m+k)}\)
12: end for

**Lemma 1:** The PDF of SNR for existing \(M\) slice-in-slice categories is \(e^{-\frac{U(m)}{N_0 b_{i,m+k}}\Delta t}\), where \(U(M)\) are UEs across \(M\) [Appendix A [18]]. Suppose a new slice-in-slice category \((k)\) is formed. The network is benefited with a throughput increment, but with the cost of more \(r_{i,m+k}\) and interference.

**Lemma 2:** The probability of formation of new slice-in-slice category conditioned on \(M\) categories per network slice is \(e^{-\frac{U(m+k)}{N_0 b_{i,m+k}}\Delta t}\), where \(\{\lambda_{1, 2}\} \Rightarrow \{\text{Appendix B [18]}\}

\[
f_d \frac{\Delta t}{M} \sum_{m=1}^{M} \sum_{i=1}^{U(m+k)} e^{\beta r_{i,m+k} s_{i,m+k}} = \sum_{m=1}^{M+k} \sum_{i=1}^{U(m+k)} e^{\beta r_{i,m+k} s_{i,m+k}}\]

(6)

1 where \(c_{M}(t) = \sum_{m=1}^{M} \sum_{i=1}^{U(m+k)} b_{i,m+k} \log(1 + \gamma_{i,m+k})\), \(c_{M+\Delta}(t) = \sum_{m=k=1}^{M+k} \sum_{i=1}^{U(m+k)} b_{i,m+k} \log(1 + \gamma_{i,m+k})\) are cell throughputs (sum of UEs service throughput (based on [12])) of \(m, m+k\) slice-in-slice categories, where \(\max(k) = \Delta\). Then, total PDF across all slice-in-slice categories is the sum of \(m\) slice-in-slice category PDF (Lemma 1) and
probability of formation of new slice-in-slice category (Lemma 2). To learn $r_{i,m+k}$, we differentiate with known $s_{i,m+k}$,
\[
\frac{\partial}{\partial s_{i,m+k}} \left( \frac{U(M)-1}{\gamma(U(M))} \right) + \frac{\partial}{\partial s_{i,m+k}} \left( e^{-\lambda_1(M)(m+k)!} \right) = 0 \Rightarrow
\]
\[
\gamma(U(M)) \left( \sum \frac{U(M)}{\gamma(U(M))} \right)_{k=0} = -FT_k
\]
\[
\{FT_k, ST_{k=0} \} = \left\{ \frac{\partial}{\partial s_{i,m+k}} \left( e^{-\lambda_2(M)(m+k)!} \right), \frac{\partial}{\partial s_{i,m+k}} \left( e^{-\lambda_1(M)(m+k)!} \right) \right\}
\]

Algorithm 2 Optimal S-TS

1. $X^n = b_{i,m+k} > 0, \epsilon, \epsilon' > 0, c_1 \in (0, 1), c_2 \in (c_1, 1)$
2. for $f_2(X^n) \leq \epsilon$ do
3. $d^n_i = -\nabla f_1(X^n)$
4. Find $\alpha^n_i$ along $d^n_i$ s.t.
5. $f_1(X^n + \alpha^n_i d^n_i) \leq f_1(X^n) + \alpha^n_i \nabla f_1(X^n) d^n_i$
6. $\nabla f_1(X^n + \alpha^n_i d^n_i) d^n_i \geq \nabla f_1(X^n) d^n_i$
7. $n \leftarrow n + 1$: $X^n = X^{n-1} + \alpha^n_i d^n_i$
8. if $|f_i(X^n) - f_i(X^{n-1})| < \epsilon'$ then
9. $b_{i,m+k} \leftarrow X^n$ \text{Eqn. (3)}
10. else
11. Go to Step 2.
12. end if
13. end for

\[
\Rightarrow \left\{ \frac{U(M)+\Delta}{\gamma(U(M))} \sum_{m,k=M+1}^M \frac{U(m+k)}{\gamma(U(M))} \right\}_{k=0} = -FT_k
\]
\[
\{FT_k, ST_{k=0} \} = \left\{ \frac{\partial}{\partial s_{i,m+k}} \left( e^{-\lambda_2(M)(m+k)!} \right), \frac{\partial}{\partial s_{i,m+k}} \left( e^{-\lambda_1(M)(m+k)!} \right) \right\}
\]

Lemma 3: The optimal weight $\beta$ is learnt, such that $O(r^*_i,m) = 0$, where $r^*_i,m = r_{i,m}(\beta)$ [See Appendix C [18]]. To solve (8) for $r_{i,m+k}$, $u_{i,m+k}(t)$, we propose a low time complexity dynamic programming Algorithm 1. As new UE services enter the network slice, $r_{i,m+k}$, $u_{i,m+k}(t)$ for previous UE services is re-learnt by re-training the model(s), to accommodate more UE services. Using the models, 

O1: Maximum additional slice-in-slice categories per network slice ($\Delta$): Consider C in (5). Solving C for $U(m) = f_a$ (base number of UE services), yields $S$. Then, $\Delta = S - m$.

O2: Maximum throughput of $\Delta$ ($\text{Thr}(\Delta)$): Let $\text{Thr}(S)$ and $\text{Thr}(m)$ be the maximum cell throughput(s) for $S$ and $m$ respectively. Then, $\text{Thr}(\Delta) = \text{Thr}(S) - \text{Thr}(m) = \sum_{m+k=1}^{M+\Delta} \sum_{i=1}^M u_{i,m+k}(t) - \sum_{m=1}^M \sum_{i=1}^M u_{i,m}(t)$. 

C. Optimal S-TS for UE(s) services

Multi-objective programming (MOP) provides several approaches to solve objectives leading to Pareto Optimality. For learning the S-TSs, i.e., $b_{i,m+k}$, we formulate the MOP

Algorithm 3 Predictive DMUSO

1: $i \leftarrow 1, 2, ..., U$ (UE services), $(N)$: Network slice 
2: $m \leftarrow 1, 2, ..., M$ (existing slice-in-slice categories)
3: $k \leftarrow 0, 1, 2, ..., N$ (additional slice-in-slice categories),
4: Initial: $m = 1, k = 0$, Scheduling $(S_1^{(N)}) = \{ i,1 \}$
5: $t \leftarrow 1, 2, ..., (T$ maximum simulation time)
6: for each TTI $t \leq T$ do
7: for $(m+k)$th category do
8: $(m+k)$th category $\leftarrow \{ i = 1, 2, ..., U(m+k) \}$
9: for $i \in (m+k)$ do
10: $S \leftarrow (m+k)$
11: $\{ r_{i,m+k}^{(N)}(t), u_{i,m+k}^{(N)}(t) \}$ model $\leftarrow$ Algorithm 1
12: $\{ r_{i,m+k}^{(N)}(t), u_{i,m+k}^{(N)}(t) \}$ $\leftarrow$ Point 10
13: $\eta_i,m+k = \sum_{i,m+k} u_{i,m+k}(t)(P_{(a)}+t_u)$ \% Sec. III-C
14: $S_1^{(N)} \leftarrow \{ (i,1), r_{i,1}, \}$ \% Updating
15: end for
16: end for
17: $m = U_{m+k} \leftarrow \sum_{i=1}^{U(m+k)} i$ \% Updating
18: for $m_{(N)} \leq M_{(N)}$ do
19: $m = (m+1)_{(N)}$ \% Updating
20: if $\sum_{m+k=M+1}^M U(m+k) r_{i,m+k} \leq r_{i,m+k}^{(N)}$ then
21: $k \leftarrow k + 1$, Go to Step 7.
22: end if
23: end for
24: $[\Delta(N) \leftarrow k_{max} \leftarrow k : S(N) = M_{(N)} + \Delta(N)]$
25: end if
26: end for
27: $i \leftarrow U_{m+k} \leftarrow \sum_{m+k=1}^{M+\Delta} i$ \% Updating
28: Go to Algorithm 4.
29: $t \leftarrow t + 1$. Go to Step 6.
30: end for

\[
\{ M_1 : max b_{i,m+k}, \eta_{i,m+k}, M_2 : max b_{i,m+k}, u_{i,m+k}(t) \} \quad (9)
\]

(a) $\eta_{i,m+k}$: From [11], $\eta_{i,m+k} = \frac{E_{i,m+k}}{P_{(a)} + t_u}$, where $E_{i,m+k}$ and $P_{(a)}$ are the energy efficiency and system circuit power. Inserting $E_{i,m+k} = \frac{u_{i,m+k}(t)}{\phi_u t_u}$ [9] yields

\[
\eta_{i,m+k} = \frac{u_{i,m+k}(t)}{P_{(a)} + \phi_u t_u(P_{(a)} + t_u)}
\]

(b) $u_{i,m+k}(t)$: From (a), inserting $\eta_{i,m+k} = \log(1+\gamma_{i,m+k})$ in (4), (9) yields

\[
M_2 : max f_d\log(1+\gamma_{i,m+k})
\]
Lemma 4: MOPs M1 and M2 are convex functions of $b_{i,m+k}$ [See Appendix D [18]]. To estimate the optimal solution, we re-express M1 and M2 as,

$$
\begin{align*}
M1: & \min & & (P_{0,i}+\phi t_o)(P_{i,m+k}+t_o) \\
& & & \log(1+\gamma_{i,m+k}) \\
& & & f_2(\delta(m+k)) \\
M2: & \min & & \Delta((P_{0,i}+\phi t_o)(P_{i,m+k}+t_o)) \\
& & & f_2(\delta(m+k)) \\
\end{align*}
$$

To solve (12), we employ the $\epsilon$-constraint and line search optimization methods, explained in Algorithm 2.

Algorithm 4 Slice-in-slice category optimal formations

1. $m + k = 1, i = 1, \delta_{\epsilon(n)} = 0.05Mbps, \delta_{\epsilon(n)} = 0.05$bits/sec/Hz (user - defined $\forall n$), $SS_i^{(N)} = \{\cdots\}$
2. for $i^{(N)} \in l_{m+k}^{(N)} \leq l \in S_i^{(N)}$ do
3. $S_i^{(N)} \leftarrow \{u, \delta_{\epsilon(n)}(\delta_{\epsilon(n)} + \lambda, u, \delta_{\epsilon(n)}(\delta_{\epsilon(n)} + \lambda)\}$ (Defn. 1: Store in $SS_i^{(N)}$)
4. $S_i^{(N)} \leftarrow i^{(N)} \leftarrow S - RS allocation$
5. $i \leftarrow i + 1$
6. if $\{u, \delta_{\epsilon(n)}(\delta_{\epsilon(n)} + \lambda)\} \in S_i^{(N)}$ then
7. $S_i^{(N)} \leftarrow i^{(N)} \leftarrow S - RS allocation$
8. else
9. $n^{(N)} \leftarrow n^{(N)} + 1$. Go to Step 3.
10. end if
11. end for
12. $(m + k)^{(N)} \leftarrow (m + k + 1)^{(N)}$. Go to Step 2.

Theorem 1: Let $f_1$ be continuous and differentiable, $X^n = b_{i,m}d_1^n \in \mathbb{R}$, such that $f_1(X^n + \alpha^i_{1}d_1^n)$ is bounded and $\nabla f_1^n d_1^n = 0$. Then, for $c_1 \in (0, 1), c_2 \in (c_1, 1), \alpha^i_n : f_1^n d_1^n = f_1^n d_1^n + c_1 f_1^n \nabla f_1 d_1^n \geq c_2 f_1^n d_1^n$ (13)

$\alpha^i_n$ is an interior in a non-empty set [See Appendix E [18]].

Theorem 2: The solution $\gamma_{i,m+k}$ as a function of $b_{i,m+k}$ is weakly pareto optimal (WPO) if $f_2(\gamma_{i,m+k}) < c_2$ or PO if $f_2(\gamma_{i,m+k}) = c_2$ [Appendix F [18]]. The complete DMUSO algorithm is shown in Algorithm 3 and 4.

IV. SIMULATION PARAMETERS

This section describes the simulation parameters set for DMUSO. The simulation comprises of n0NodeBs serving randomly distributed UE(s) services. For simulation we consider 4 network slices of bandwidth 40, 60, 70 and 80 MHz respectively. Each UE hosts 4 services across 4 network slices which are optimally grouped into slice-in-slice categories. The detailed system-level simulation parameters are in Table I.

| Parameter | Value |
|-----------|-------|
| Channel bandwidth | 100 MHz |
| Network slice(s) topology | Four network slices with $M = 6$, Each UE serves 4 applications across 4 network slices. Each slice-in-slice category hosts 5 UE service(s) initially. |
| Bandwidth part (BP): Maximum S-TSs | 40 MHz ($N_1$), 60 MHz ($N_2$), 70 MHz ($N_3$) |
| Maximum S-RSs | 200 ($N_1$), 300 ($N_2$), 350 ($N_3$), 400 ($N_4$) |
| Total number of UEs associated with network slices | 205 ($N_1$), 335 ($N_2$), 385 ($N_3$), 455 ($N_4$) |
| Channel models | Extended pedestrian A model 5 Hz |
| UE mobility | 5 m/sec - 35 m/sec |
| UE transmit power | 25 dBm |
| Cell radius | 95 km |
| Modulation | QPSK for $-\infty < \text{SNR} < 8$ dB |
| | 16-QAM for $8 \text{ dB} < \text{SNR} < 14$ dB |
| | 64-QAM for $14 \text{ dB} < \text{SNR} < \infty$ |
| Simulation time | 100 seconds (10000 TTI) |
| Simulators | NetSim and MATLAB |

Fig. 2(a) shows the interference function $(\log(1 + \text{SINR}))$ vs optimal S-TSs across UE services. From Fig. 2 (a), $\log(1 + \text{SINR})$ are weakly PO for range of S-TSs along X-axis. This is because along the Y-axis, the interference function increases as S-TSs increases, depicting strong PO. From the concept of strong PO [11], any strong PO is weakly PO $\Rightarrow \exists! b_{i,m+k} \rightarrow \log(1 + \text{SINR})_{b_{i,m+k}} \leq \log(1 + \text{SINR})_{b_{i,m+k}} \rightarrow \text{PO}$. As the interference function decreases (Y-axis of Fig. 2(a)), at lower S-TSs, $\log(1 + \text{SINR})_{b_{i,m+k}} \approx \log(1 + \text{SINR})_{b_{i,m+k}}$ for $b_{i,m+k} \approx b_{i,m+k}$, depicting weakly PO. DMUSO uses the S-TSs in learning the optimal S-RSs allocation, shown in Fig. 2(b). From Fig. 2(b), for $N_1, N_2, N_3, N_4$, as compared to the maximum S-RSs (Table I), DMUSO performs near (to maximum S-RSs) and optimized S-RS allocation across UE services. The optimal S-RSs estimates the UE service throughput. The sum of the throughputs across UE services over TTIs gives the average cell throughput. Fig. 3 shows the weighted mean average cell throughput (Mbps) across 10000 TTIs vs slice-in-slice categories served across $N_1, N_2, N_3, N_4$. From the X-Y plane of Fig. 3, the maximum slice-in-slice categories S (including $M$ (Table I)) served by $N_1, N_2, N_3, N_4$ are {41, 67, 77, 91} owing to the maximum S-RSs resource constraint. As the number of slice-in-slice categories increases, average cell throughput increases across the network slices. A clearer view of the performance of initial few slice-in-slice categories can be seen in the zoom portion of Fig. 3. In addition, as compared to state-of-the-art algorithms, DMUSO ensures minimum and maximum gains of 4.5 and 7.5. The performance improvement is because DMUSO estimates the dynamic optimum S-RSs and throughput for each UE service, which is re-learnt to accommodate more UE services within network slices. We now analyze the average cell throughput as a function of SNR and S-RSs. We plot the performances for the network slices having minimum and maximum BP, i.e., $N_1$ and $N_4$. Fig. 4 shows the average cell
We proposed novel predictive Dynamic Scaling Multi-slice-in-slice-connected UEs services for system resource optimized Scheduling (DMUSO) models and algorithms. Analysing the results, DMUSO achieves efficient and optimized system resource scheduling with significant performance gains of atleast 4.4 and 7.5 times compared to state-of-the-art algorithms. As part of future work, we intend to analyze the effect of varying UE mobility on resource scheduling across network slices.

ACKNOWLEDGMENT

The authors would like to thank Advanced Communication System project sponsored by NM-ICPS scheme of Department of Science and Technology, Govt. of India.

REFERENCES

[1] O-RAN WG 1 Slicing Architecture, v01.00, Technical Specification.
[2] M. Motalleb, V. Mansouri, S. Naghadeh, Joint Power Allocation and Network Slicing in an Open RAN System, in arXiv:1911.07904, Computer Science, Network and Internet Architecture, Nov. 2019.
[3] S. Chaudhuri, I. Baig, Debabrata Das, QoS aware Downlink Scheduler for a Carrier Aggregation LTE-Advance Network with Efficient Carrier Power Control, in IEEE INFOCOM, Bangalore, India, 2018.
[4] L. Zhou, T. Zhang, J. Li, Y. Zhu, Radio Resource Allocation for RAN Slicing in Mobile Networks, Chongqing, China, IEEE/CIC International Conference on Communications in China, Aug. 2020.
[5] 3GPP TS 38.300 v15.2.0 Release 15: 5G NR Overall description-2.
[6] Y. Saito, et. al., Flexible Resource Block Allocation to Multiple Slices for Radio Access Network Slicing Using Deep Reinforcement Learning, IEEE Access, vol. 8, pp. 68183-68198, Apr. 2020.
[7] R. Stoica, G. Abreu, 6G: the Wireless Communications Network for Collaborative and AI Applications, in arXiv. Apr. 2019.
[8] O.U.Akgul, I. Malanchini, A. Capone, Dynamic Resource Trading in Sliced Mobile Networks, in IEEE Transactions on Network and Service Management, pp. 220-233, vol. 16, no. 1, Mar. 2019.
[9] S. Ravindran, S. Chaudhuri, J. Bapat, and Debabrata Das, EESO: Energy Efficient System-resource Optimization of Multi-Sub-Slice-Connected User in 5G RAN, in IEEE International Conference on Electronics, Communication and Computing Technologies, Bangalore, India, 2020.
[10] ORAN-WG2.AIML: AI/ML workflow description and requirements.
[11] L. Shou et. al., A New Relation Between Energy Efficiency and Spectral Efficiency in Wireless Communications Systems, in IEEE Wireless Communications, vol. 26, no. 3, pp. 168-174, Jun. 2019.
[12] B. Matthiesen, O. Aydin, E. A. Jorswieck, Throughput and Energy-Efficient Network Slicing, in 22nd International ITG Workshop on Smart Antennas, Bochum, Germany, 2018.
[13] M. Yan, G. Feng, J. Zhou, Y. Sun, Y. Liang, Intelligent Resource Scheduling for 5G Radio Access Network Slicing, in IEEE Transactions on Vehicular Technology, vol. 68, no. 8, Aug. 2019, pp. 7691 - 7703.
[14] F. Fossati, et. al., Multi-Resource Allocation for Network Slicing, in IEEE/ACM Transactions on Networking (Early Access), Mar. 2020.
[15] M. Dighriri et. al., Resource Allocation Scheme in 5G Network Slices, in 32nd International Conference on Advanced Information Networking and Applications Workshops, Krakow, Poland, May 2018.
[16] Y. Luo, X. Ye, X. Wang, H. Miao, T. Yang, An Optimal Scheduling Strategy for Wireless Resources based on Network Slicing, in IEEE 3rd Advanced Information Management, Communicates, Electronic and Automation Control Conference, Chongqing, China, Oct. 2019.
[17] H. Ko, J. Lee, S. Pack, Priority-Based Dynamic Resource Allocation Scheme in Network Slicing, 2021 International Conference on Information Networking, Jeju Island, Korea (South), Jan. 2021.
[18] Proofs of Lemmas and Theorems, Available [Online]: https://bit.ly/3dw9KpM