QoS-aware Resource Management with Network Slicing and Beam Management

Amber Hina
DET, Politecnico di Torino
Torino, Italy

Daniele Brevi
LINKS Foundation
Torino, Italy

Carla Fabiana Chiasserini
DET, Politecnico di Torino
Torino, Italy

Alessandro Nordio
CNR-IEIIT
Torino, Italy

Claudio Pastrone
LINKS Foundation
Torino, Italy

Alberto Tarable
CNR-IEIIT
Torino, Italy

***Abstract*—Beyond-5G and 6G networks must support diverse QoS, from ultra-reliable low latency communications to enhanced mobile broadband. This requires joint, efficient resource management that combines network slicing and dynamic beam management. This paper addresses the complexity of optimizing resource allocation across users and slices served by a set of beams. We propose SBQF, a high-performance technique that optimizes resource allocation at both the MAC and the physical layers. SBQF is designed to maximize the overall user utility and is in particular efficient in improving the performance of users requiring latency-critical services. Validated using real-world vehicular traces, SBQF significantly improves latency, user utility, and data rates over existing solutions.**

***Index Terms*—Qos, resource management, fractional programming, beamforming, network slicing.**

I. Introduction

Modern wireless networks, such as beyond-5G and 6G (B5G/6G), are designed to support a wide range of users with different service demands. They must meet various quality of service (QoS) requirements, ranging from ultra-reliable low-latency communications for real-time and mission-critical applications to massive machine-type communications and enhanced mobile broadband for data-intensive applications. In order to meet these demands, the idea of network slicing (NS) has emerged as a crucial enabling technology for B5G/6G networks, allowing the setup of multiple virtual networks atop a single physical infrastructure. It ensures that different service level agreements (SLAs) associated with each slice are satisfied, such as high data rate, low latency, and high reliability. Network slicing distributes resources across different classes of users in an efficient manner; however, it adds more complexity in efficient radio resource management. At the physical layer, antenna arrays play an important role, as they generate directional beams focusing radio signals towards intended users, so as to reduce interference and increasing data rates. Since beam orientation, beamwidth, and transmission power can be dynamically adjusted in real time to adapt to variations in network topology and coverage demands, efficient beam management plays a vital role in enhancing overall network performance.

Network slicing and beam management have received significant attention in recent years. Specifically, the topic of cost-effective network slicing for eMBB and uRLLC traffic has been addressed in [1]. Of particular relevance to our work is the study on 5G NR radio resource management in [2], where the author proposed a radio resource management approach featuring network slicing and handling both latency and throughput requirements.

The work in [3] studies the problem of allocating resources in sliced networks and proposes a simple dynamic resource sharing policy. Network slicing and beam management have been jointly explored in [4], which introduces an efficient beam optimization scheme for multiple slices using a deep reinforcement learning framework. The study in [5] optimizes radio resource allocation in 5G networks to satisfy QoS requirements while reducing overall resource consumption.

This paper addresses the complex problem of optimizing resource allocation in a sliced network where users are served by directional beams, with the goal of maximizing the users' QoS. Unlike prior studies, this work jointly optimizes resource allocation at physical and MAC layers to guarantee the target service demands of various user types.

The primary contributions of this work are as follows:

- We develop a dynamic network model supporting the coexistence of users with different QoS requirements through slices. The model captures important network features at physical and MAC layers, such as interference mitigation, radio resource allocation among users, and data packet generation.
- We propose an efficient and high performance radio resource allocation technique, named *Slice-beam Queue Fractional programming* (SBQF) maximizing the users' utility. Such technique provides significant performance gain over state-of-art solutions. SBQF is particularly focused on minimizing the latency of users requiring low-latency network services.
- We validate our approach using real-world vehicular mobility traces and demonstrate the gain that SBQF can achieve in terms of latency, user utility and data rate.

ISBN: 978-3-903176-79-9

II. System Model and Problem Statement

We consider a wireless network consisting of a set of G base stations (BSs), $\mathcal{G}$, serving a set of users, $\mathcal{U}$, which generate different traffic types and require mobile services with different characteristics. In order to accommodate such requirements, the network partitions the available radio resource into a set of slices $\mathcal{S}$, each customized to a specific mobile service. Slice $i \in \mathcal{S}$ is allocated a share $s_i > 0$, which denotes the fraction of total network resources assigned to slice i, with $\sum_{i\in\mathcal{S}} s_i = 1$. The resources assigned to slice i are then distributed among the users belonging to it. We specifically consider uplink communications and two types of mobile services: (i) constant bit rate (CBR) traffic characterized by a minimum data rate, such as that generated by video conferencing applications ($i = 1$) and (ii) variable bit rate traffic (VBR), generated e.g. by IoT devices, typically bursty and latency-sensitive ($i = 2$).

IoT devices typically follow an on/off behavior, which we model as a discrete-time stochastic process driven by a two-state Markov chain with time-step Δ_t. When in the ”on” state, at each time step an IoT user generates data packets of size C bits, according to a Poisson distribution with parameter λ. Each packet is marked with its generation time, t_{g}, and is placed in a queue, waiting to be transmitted. VBR users in the ”off” state remain silent and do not generate traffic.

The network area is divided into zones, whose set is denoted by $\mathcal{Z}$. At any given time t, zone $z \in \mathcal{Z}$ consists of a set of users $\mathcal{U}_z(t)$, such that $\cup_{z\in\mathcal{Z}}\mathcal{U}_z(t) = \mathcal{U}(t)$, which may include CBR and VBR user types. The set of users of type i in zone z is denoted as $\mathcal{U}_{z,i}(t)$. Also, at time t user u is assigned to exactly one zone, i.e, $\mathcal{U}_{z,i}(t) \cap \mathcal{U}_{z',i}(t) = \emptyset$, for all $z \neq z'$ and $i \in \mathcal{S}$. For simplicity, from now on the time dependency is dropped from the notation.

We assume that the BSs are identical to each other and operate on the same frequency band, with bandwidth B. Through antenna arrays they can generate narrow beam focusing enhancing the communication towards subsets of users.

Specifically, BS $g \in \mathcal{G}$ serves the set of zones $\mathcal{Z}_g$, using $|\mathcal{Z}_g|$ beams, i.e., one beam per zone. Users’ are assumed to be equipped with omnidirectional antennas. The power radiated by users u is denoted by p_u.

Radio resources can be modeled as a set of discrete time-frequency blocks, $\mathcal{F}$. Each user, u, is assigned a subset $\mathcal{F}_u \subseteq \mathcal{F}$ of them, whose measure is $f_u = |\mathcal{F}_u|/|\mathcal{F}|$, which represents the fraction of resources assigned to a user by its associated BS. Typically, users associated with the same BS are assigned non-overlapping subsets of resources, i.e., they do not interfere with each other. Also, the allocated resources should satisfy the constraints

$$\sum_{z\in\mathcal{Z}_g}\sum_{u\in\mathcal{U}_z} f_u = 1 \quad \forall g \in \mathcal{G} \tag{1}$$

$$\sum_{z\in\mathcal{Z}}\sum_{u\in\mathcal{U}_{z,i}} f_u \leq G \cdot s_i \quad \forall i \in \mathcal{S} \tag{2}$$

which ensure that all available radio resources are distributed among users and that the constraints imposed by slicing, are met.

Furthermore, the channel connecting user u to BS g through the beam serving zone $z \in \mathcal{Z}_g$ is denoted by the random complex scalar coefficient $h_{z,g,u}$.

A. Performance metrics

As a performance metric for the scenario under investigation, we consider the generalized network utility function [6]

$$U = \sum_{i\in\mathcal{S}} U_i \tag{3}$$

where $U_i = \frac{1}{|\mathcal{U}_i|}\sum_{u\in\mathcal{U}_i}\operatorname{erf}(Q_u)$ is the utility achieved by users of slice i. In the above expression the term Q_u represents the QoS metric associated to user u, given by

$$Q_u = \begin{cases} r_u/r_u^*, & \text{if } u \text{ generates CBR traffic,} \\ l_u^*/l_u, & \text{if } u \text{ generates VBR traffic,} \end{cases} \tag{4}$$

where r_u is the data rate achieved by user u, r_u^* is the target data rate, l_u is the achieved latency, and l_u^* is the target latency. The achieved rate r_u in (4) is computed as

$$r_u = Bf_u \log_2\left(1 + \mathrm{SINR}_u\right) \tag{5}$$

where SINR_u is the SINR of the signal transmitted by user $u \in \mathcal{U}_z$ and received by its serving BS g, given by

$$\mathrm{SINR}_u = \frac{p_u|h_{z,g,u}|^2}{\displaystyle\sum_{\substack{g'\in\mathcal{G}\\ g'\neq g}}\sum_{z'\in\mathcal{Z}_{g'}}\sum_{u'\in\mathcal{U}_{z'}} p_{u'}|h_{z,g,u'}|^2\frac{|\mathcal{F}_u \cap \mathcal{F}_{u'}|}{f_{u'}} + N_0 B f_u}\,. \tag{6}$$

The network sum-rate is defined as $R = \sum_{i\in\mathcal{S}} R_i$ where $R_i = \sum_{u\in\mathcal{U}_i} r_u$ is the sum rate achieved by users in slice i. In (6) N_0 is the thermal noise power spectral density and the term $|\mathcal{F}_u \cap \mathcal{F}_{u'}|$ represents the overlap of resources allocated to users u and u' which is proportional to the interference that user u experiences from user u'.

For each packet transmitted by VBR user u, the latency l_u is computed as

$$l_u = t_{\mathrm{r,u}} - t_{\mathrm{g,u}} \tag{7}$$

where $t_{\mathrm{g,u}}$ is the packet generation time and $t_{\mathrm{r,u}}$ is the packet successful transmission time. As previously mentioned, VBR users maintain a queue where generated packets are placed and wait to be transmitted. At any given time, the length of such queue is denoted by L_u which indicates the number of packets awaiting service. For VBR users, the number of packets that can be transmitted in a given time interval depends on r_u. By increasing r_u more packets can be transmitted thus reducing their waiting time and, hence, their latency. Given the above described network model we aim at maximizing the network utility, U, by solving the following optimization problem:

$$\max_{\{\mathcal{F}_u\}} \quad U \qquad \text{s.t. (1), and (2)}\,. \tag{8}$$

This problem presents significant challenges since finding the optimal sets of resources $\{F_u\}_{u\in\mathcal{U}}$ satisfying the constraints (1) and (2) is a combinatorial problem whose complexity increases with the number of users $|\mathcal{U}|$ and with the total number of time-frequency blocks in the resource space, which are typically large numbers.

III. SBQF:The SliceBeam-QF solution framework

In light of the complexity of solving problem (8), we propose to tackle it in two steps:

- we first compute the per-user resource allocations, $\{f_u\}_{u\in\mathcal{U}}$, using the share-constrained proportionally fair (SCPF) technique proposed in [3]. SCPF satisfies constraints (1) and (2) and, hence will be considered as a baseline approach. We will then enhance it by proposing an allocation based on the queues length L_u so as to improve the performance of VBR users and to reduce the latency they experience.
- given the set $\{f_u\}_{u\in\mathcal{U}}$, we maximize the network sum-rate R, by minimizing the interference among BSs. This is achieved by implementing an optimization technique based on fractional programming (FP).

A. Queue-based resource allocation

The SCPF allocation scheme distributes the slice share equally among its users. This is obtained by first computing the weights $w_u = s_i/|\mathcal{U}_i|$ for all $u \in \mathcal{U}_i$ and $i \in \mathcal{S}$. Then, each base station allocates resources to users in proportion to their weights, irrespectively of the users' state and of the time evolution of their queue lengths. This *static allocation* (SA) of resources is obtained by SCPF by computing for user u served by BS g

$$f_u^{(\mathrm{SA})} = \frac{w_u}{\sum_{z\in\mathcal{Z}_g}\sum_{v\in\mathcal{U}_z} w_v}\,. \tag{9}$$

Since latency can be mitigated by reducing the waiting time of the packets in the users' queues, we propose to assign resources to users proportionally to their queue length. According to this *queue based allocation* (QBA), the resource shares are computed as

$$f_u^{(\mathrm{QBA})} = \frac{f_{z,2}\,L_u}{\sum_{v\in\widehat{\mathcal{U}}_{z,2}} L_v}, \quad \forall\, u \in \widehat{\mathcal{U}}_{z,2} \tag{10}$$

where $\widehat{\mathcal{U}}_{z,2}$ is the set of VBR users in zone z having a non-zero queue length and $f_{z,2} = \sum_{u\in\mathcal{U}_{z,2}} f_u^{(\mathrm{SA})}$.

B. Interference Minimization through FP

The QBA technique described in Sec. III-A only provides a way to compute the measures $\{f_u\}_{u\in\mathcal{U}}$ of the sets $\{\mathcal{F}_u\}_{u\in\mathcal{U}}$. However, the $\mathcal{F}_u$'s play a key role in the SINR expression (6) since they affect the amount of interference experienced by the users and, hence, the achieved SINR. Finding the optimal sets $\{\mathcal{F}_u\}_{u\in\mathcal{U}}$ that maximize some performance metric, e.g. the network sum-rate R_s, given $\{f_u\}_{u\in\mathcal{U}}$, is a combinatorial optimization problem. Due to its complexity, brute-force solutions are practical only for small-scale scenarios [7]. The FP technique [8], is a viable approach for handling this problem with an acceptable complexity. FP takes as input the set of channel coefficients $\{h_{z,g,u}\}_{z\in\mathcal{Z},g\in\mathcal{G},u\in\mathcal{U}}$, and the set of resources $\{f_u\}_{u\in\mathcal{U}}$, providing as output the optimal value for the decision variables $\beta_{u_1,\dots,u_G} \geq 0$ denoting the amount of resources contemporarily assigned to users $u_1,\dots,u_G$ where $u_g \in \mathcal{U}^{(g)}$, $g = 1,\dots,G$ and $\mathcal{U}^{(g)}$ is the set of users associated to BS g. Then, the SINR achieved by user u, $\mathrm{SINR}_u^{(\mathrm{FP})}$, can be computed using (6) where the term $|\mathcal{F}_u \cap \mathcal{F}_{u'}|$ is replaced with $\gamma_{u,u',g,g'}$ defined as

$$\gamma_{u,u',g,g'} = \sum_{\mathbf{a}\in\mathcal{A}_{u,u',g,g'}} \beta_{a_1,\dots,a_G}$$

where $\mathcal{A}_{u,u',g,g'} = \{\mathbf{a} \in \mathcal{A} | a_g = u, a_{g'} = u'\}$ and $\mathcal{A} = \{\mathbf{a} = (a_1,\dots,a_G) | a_1 \in \mathcal{U}^{(1)},\dots,a_G \in \mathcal{U}^{(G)}\}$.

IV. Numerical results

We consider real-world mobility traces for the city of Cologne (Germany). A portion of the city comprising $G = 6$ BSs is divided into zones of size 200 m as described in [7]. BSs are equipped with a 16-elements linear array of antennas each providing 8 dBi gain while user antennas are isotropic. The transmit power of all user equipment is set to P_u=23 dBm, for all $u \in \mathcal{U}$; the carrier frequency is 3 GHz and the bandwidth $B = 50$ MHz. The channel coefficients $h_{z,g,u}$ are drawn from the urban micro street canyon model [9]. At any given time, the mobility model provides the number of vehicles falling in each zone but does not differentiate between users requiring different service types. Therefore, for each zone we set $|\mathcal{U}_{z,2}| = 2|\mathcal{U}_{z,1}|$. Also, we assume that CBR users are using video conferencing applications with minimum data rate $r_u^* = 4$ Mb/s, while VBR users are IoT devices that allow a maximum target latency $l_u^* = 3$ ms. For the corresponding two network slices we set the shares $s_1 = 0.7$ and $s_2 = 0.3$. IoT users generate packets with size $C = 2000$ bits according to a discrete-time Poisson process with arrival rate $\lambda = 10$ and time-step $\Delta_t = 10$ ms.

At each time-step Δ_t, (i) the queues of the IoT users are updated by adding the newly generated packets; (ii) new instances of the channels coefficients $h_{z,g,u}$ are drawn; (iii) the shares $\{f_u\}$ are re-computed according to (10), (iv) FP is run to maximize the sum rate R and, finally, (v) some packets waiting in the queues are transmitted according to the rate achieved by the IoT users.

Figure 1 shows the cumulative density function (CDF) of the network utility, U, and of the IoT users' utility U_2, defined in (3) for three different resource allocation techniques. The blue curve represents the baseline case, where resources are assigned to IoT users through SA, given by (9) and the sets $\{\mathcal{F}_u\}$ have been randomly assigned (RA) in the radio resource space $\mathcal{F}$ to provide average network performance. We call this configuration *static allocation random assignment* (SARA).

The green curve represents the case where the sets $\{\mathcal{F}_u\}$ have been randomly chosen and in each zone the resources are assigned only to the set of IoT users, $\bar{\mathcal{U}}_{z,2}$, being in "on" state or having packets in their queues waiting to be transmitted. Accordingly, the resource are computed as $f_u^{(\mathrm{DA})} = \frac{f_{z,2}}{|\bar{\mathcal{U}}_{z,2}|}$, for all $u \in \bar{\mathcal{U}}_{z,2}$. We call this approach *dynamic allocation random assignment* (DARA) since it adapts to the evolution in time of

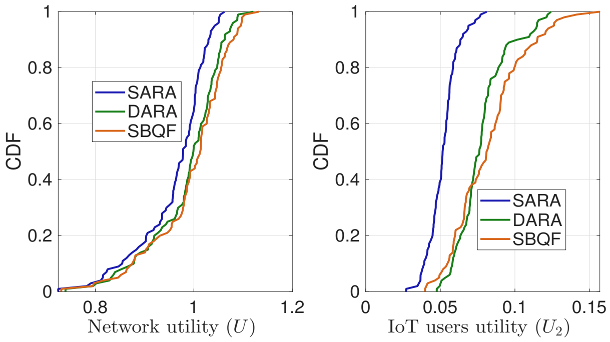

Fig. 1. CDF of Network utility (U) and IoT users utility (U_2)

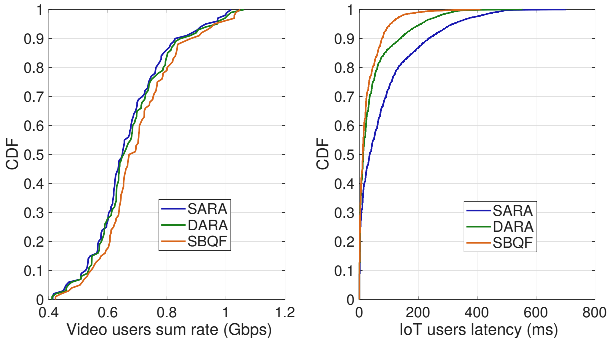

Fig. 2. Sum rate for the video users and latency for IoT users

the users' states, but does not take into account for the queues length. Finally, the orange curve shows the performance of our proposed SBQF approach which uses FP and (10). From Fig. 1(left) we observe that DARA provides a utility gain w.r.t. the baseline case showing the benefit that a dynamic allocation of the resources can have on the entire network. However, if the resource allocation takes also into account for the queues length and the SINR is optimized through FP a further gain can be achieved. The advantage of implementing SBQF has clearly more impact on the performance of the IoT users, as can be observed in Fig. 1(right). In this case, the SBQF gain w.r.t. to SARA is significant. With SBQF, the probability of the IoT utility dropping below $U_2 = 0.08$ reduces to 47%, compared to 68% for DARA and 100% SARA. Figure 2(left) illustrates the CDF of the sum rate achieved by video users, R_1 where SARA and DARA exhibit comparable sum rates. However, SBQF outperforms all of them by providing 7.5% gain w.r.t. to SARA and 6% improvement over DARA. The reason for this gain is mainly due to FP, since the resource allocation according to the queue lengths in (10) only affects IoT users.

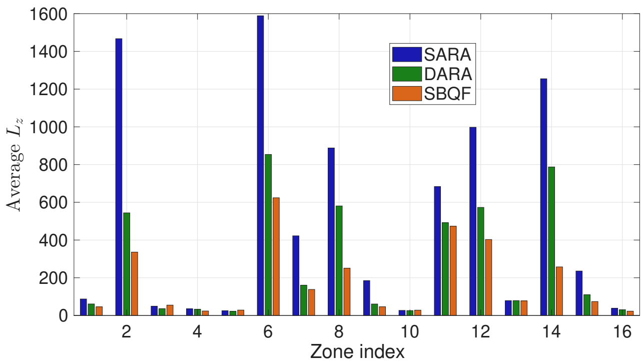

Fig. 3. Average queue length per zone

Fig. 2(right), shows the CDF of the latency experienced by IoT users. Here the latency reduction provided by SBQF is substantial. As an example, the probability of showing latencies larger than 60 ms is about 17% while it raises to about 40% for SARA. These results indicate that, although FP is crucial for maximizing data rates for video users, latency minimization mainly depends on the implementation of effective resource allocation strategies.

Fig. 3 depicts the average over time of the queue length per zone, defined as $L_z = \sum_{u \in \mathcal{U}_{z,2}} L_u$, for SARA, DARA and SBQF. A large average L_z can be due to the zone having a large number of IoT users or experiencing very low SINR that prevents the transmission of a large number of packets per time step. In any case SBQF can reduce the average queue length up to a factor 4 w.r.t. SARA, and up to a factor 2 w.r.t. DARA, depending on the zone.

V. Conclusion

In this study, we explored network slicing and beam management to enhance overall network performance. We developed a novel solution framework, named SBQF, that integrates queue-based resource allocation with a fractional-programming technique to optimize both radio and physical layer resources. Simulation results indicate that SBQF outperforms baseline schemes by achieving higher data rates and reduced latency across various service slices. These findings demonstrate the potential of proposed approach to provide efficient and flexible resource management under dynamic network conditions.

References

[1] S. Pramanik, A. Ksentini, and C. F. Chiasserini, "Cost-efficient RAN slicing for service provisioning in 5G/B5G," *Computer Communications*, vol. 222, pp. 141–149, 2024.

[2] K. Boutiba, M. Bagaa, and A. Ksentini, "Optimal radio resource management in 5G NR featuring network slicing," *Computer Networks*, vol. 234, p. 109937, 2023.

[3] J. Zheng, P. Caballero, G. de Veciana, S. J. Baek, and A. Banchs, "Statistical multiplexing and traffic shaping games for network slicing," *IEEE/ACM Trans. on Networking*, vol. 26, no. 6, pp. 2528–2541, 2018.

[4] O. Sabr, G. Kaddoum, and K. Kaur, "PABSO-DRL: Power and beam self-optimization scheme for multiple slices in MU-MISO systems," *IEEE Transactions on Consumer Electronics*, pp. 1–1, 2024.

[5] N. Villegas, J. L. Herrera, L. Diez, D. Scotece, L. Foschini, and R. Agüero, "DRL-based dynamic mac scheduler reconfiguration in O-RAN," in *ICC 2025*, 2025, pp. 5023–5028.

[6] T. Hu, Q. Liao, Q. Liu, A. Massaro, and G. Carle, "Fast and scalable network slicing by integrating deep learning with lagrangian methods," in *GLOBECOM*, 2023, pp. 6346–6351.

[7] A. Hina, D. Brevi, C. F. Chiasserini, A. Nordio, C. Pastrone, and A. Tarable, "Optimizing resource allocation for resilient networks: A beam management and network slicing approach," in *Proc. European Wireless Conference*, 2025.

[8] K. Shen and W. Yu, "Fractional programming for communication systems—part i: Power control and beamforming," *IEEE Transactions on Signal Processing*, vol. 66, no. 10, pp. 2616–2630, 2018.

[9] ETSI, "5G; study on channel model for frequencies from 0.5 to 100 GHz (3GPP TR 38.901 Release 16)," Tech. Rep., Nov. 2020.