Spectrally Efficient Guaranteed Rate Scheduling for Heterogeneous QoS Constrained Wireless Networks

Geetha Chandrasekaran, Gustavo de Veciana
Dept. ECE, The University of Texas at Austin
(geethac, deveciana)@utexas.edu

Vishnu Ratnam, Hao Chen and Charlie Zhang
Samsung Research America
(vishnu.r, hao.chen1 and jianzhong.z)@samsung.com

***Abstract*—Next generation wireless schedulers will support increasingly heterogeneous users/devices in terms of their traffic characteristics and service requirements. Particularly challenging is the need to deliver low latency traffic with strict deadlines in a spectrally efficient manner. We introduce a class of wireless schedulers, Opportunistic Guaranteed Rate (OGRS) that exploits the temporal variability in users' channel capacity with a view on maintaining delay guarantees. OGRS meets the user's delay constraints by opportunistically allocating the user the equivalent of a fixed service rate, which given a dual leaky bucket constraint on its traffic will ensure the delay requirements are met. We consider offline policies with access to future channel rates, which establishes a bound to the wireless spectral efficiency. We show via extensive simulations that OGRS can be within 10%-40% of this bound for a range of delays that were considered. These gains translate to more than a two fold enhancement in eMBB users' throughput, when URLLC and eMBB traffic share resources. Finally, we propose a measurement based admission control strategy for latency constrained URLLC users, so that the network can guarantee QoS to all its users - existing as well as newly admitted ones.**

***Index Terms*—Delay deadline, Opportunistic scheduling, Low Latency, QoS constraints, Leaky bucket, URLLC scheduling.**

I. Introduction

THE support of Ultra Reliable Low Latency Communication (URLLC) is expected to be critical towards enabling next generation [1] wireless applications such as industrial automation, augmented and virtual reality, autonomous driving, remote diagnosis and health care. The key challenge in supporting such applications is their stringent constraints on Quality of Service (QoS). The latency constraints for these applications range between 5 and 30 ms, with reliability requirements of 99.9 to 99.9999%, see e.g., [2]. Moreover, given the limited spectrum available and associated costs, it is also critical to deliver such URLLC based services in a spectrally efficient manner. In general, this is challenging, e.g., one must add substantial upfront redundancy to meet reliability requirements without delays associated with re-transmissions, or given low latency requirements one may not be able to exploit opportunism or wait for data to achieve more efficient modes of transmission.

In addition to dealing with the requirements of URLLC traffic, it is also critical to devise resource allocation and scheduling strategies that enable the support, of a mix of traffic, e.g., Enhanced Mobile Broadband (eMBB) and Machine-Type Communications (MTC) traffic, and possibly network slices provisioned to support different classes of applications. Our focus in this paper will be on spectrally efficient scheduling of wireless user traffic with possibly heterogeneous delay deadlines, perhaps the most challenging traffic class, yet we aim to provide an approach that can be combined with other scheduling policies, e.g., proportionally fair or utility maximizing schedulers used to support eMBB traffic, to manage an assortment of services with diverse QoS requirements. Below we provide a brief summary of related work in this area, focused primarily on scheduling with delay based QoS constraints. We then introduce the key contributions of this paper.

A. Related Work

Wireless scheduling can be based on the user's queue length, channel quality, history of past allocations, etc., and may have multiple objectives including Quality of Service (QoS) and fairness. In settings where users' queues are fully backlogged, perhaps the best known strategies are utility maximizing, i.e., maximizing the sum of users' utilities, which in turn is a function of each user's long term throughput. Perhaps the most popular wireless scheduling often used in practice is the proportionally fair scheduler, see [3], which maximizes the log utility of users' long term throughput. Such an approach results in users to *fair* long term rates with *opportunistic* scheduling in the short term when the channel rate is high. A more sensitive resource allocation that averts short term neglect of user allocation, see [4], maximizes the user's utility which is a function of the short term throughput. In general, such utility maximizing schedulers are best suited for elastic traffic with no hard deadlines. Several questions such as how to choose the fairness criterion when user queues are not fully backlogged or when there are reliability constraints for meeting strict delays remain unanswered.

In settings where user queues are not fully backlogged but instead driven by stochastic arrivals, various channel state dependent throughput optimal policies (that guarantee queue stability when feasible) have been devised, e.g., [5], and may also achieve different types of delay objectives, e.g., roughly minimizing the max delays across users or overall average delays. While efficient for "best effort" type traffic, such scheduling policies do not deliver strict delay guarantees needed for real-time applications and/or URLLC based services. Some of these schedulers also address other

ISBN 978-3-903176-55-3 © 2023 IFIP

performance objectives, such as minimizing the max user queue length in [6], or minimizing the average delay as in [7]. Adaptations of throughput optimal schedulers to practical settings, like Modified Largest Weighted Delay First (MLWDF) [8] consider the channel state, head-of-line packet delays and user weights reflecting QoS objectives for scheduling. Such schedulers offer a graceful degradation of service when there are insufficient resources to meet QoS of all users.

Another interesting line of research borrows ideas from wireline scheduling (e.g., traffic shaping and network calculus [9], [10]) to satisfy user QoS constraints under wireless channel variations. Weighted round robin [11] or weighted fair queueing [12] employ heuristic user weights or tokens [13] based on service deficit [14] to either minimize the average delay or provide a graceful degradation of service. Much of the above mentioned work focuses on scheduling one class of users, or traffic that is sensitive to packet delays. In practice, wireless systems need to be shared by heterogeneous user classes. Packet level deficit tracking for evaluating the QoS service deficit has been considered for each user in [15], however, such an approach is prohibitively expensive in complexity when there are a large number of users. In contrast, we employ cumulative service based techniques and queue based scheduling, avoiding the need to track packet level deadlines or control. Wireless scheduling for optimizing both service regularity and mean delay is considered in [16], but the emphasis is on graceful degradation rather than guaranteed latency. Although [17] considers scheduling with reliability for homogeneous user QoS requirements, it is assumed that only one user can be scheduled every time slot which is a severe limitation under practical scenarios.

Scheduling with guaranteed QoS is considered in [18], with no improvement in spectral efficiency for latency constrained users. Joint resource allocation for URLLC and eMBB traffic is proposed in [19] but opportunistic scheduling is limited to eMBB users. More recent literature on URLLC scheduling [20]–[24] include reliability guarantees, however, opportunistic scheduling has not been given much consideration apart from the perspective of energy efficiency [25] or the violation of deadline probability [26] or devising token based quality assurance [13] which may starve weaker users until one is forced to schedule close to their deadline. Previous research can only be considered a first step towards a more profound understanding of developing spectrally efficient algorithms for delay constrained traffic. To the best of our knowledge, a simple approach to opportunistic scheduling over temporal channel variations for deadline constrained traffic has not been considered in the existing literature. In this work, we propose a new class of opportunistic scheduling algorithm for URLLC users with heterogeneous traffic and disparate QoS requirements, to enhance the throughput/utility for eMBB traffic that also share the overall network resources.

Studies on QoS provisioning cannot be considered complete without addressing the question of admission control and/or traffic shaping/policing. A closer look at existing literature reveals that much of the work on wireless scheduling does not solve this problem. Given the wireless channel uncertainty, it is infeasible to predict if a given scheduling policy will be able to meet the user's QoS constraints with high reliability. Given the uncertainty and heterogeneity associated with traffic, channels, and user requirements in a wireless system, it is virtually impossible to devise good models that would allow one to predict if the users' QoS requirements will be met under a given scheduling policy. While there have been many works in the literature on Measurement Based Admission Control (MBAC) [27]–[29], it has to be noted that none of those accurately meet the packet loss targets [30] under finite buffer sizes, which is sufficiently similar to delay violation probability.

B. Our contributions

We propose a class of wireless schedulers that under appropriate assumptions can meet heterogeneous delay deadlines and do so in a spectrally efficient manner such that the more relaxed the constraint to more efficient. The key underlying idea is to leverage the flexibility of wireless systems, in terms of allocating a time varying number of Resource Blocks (RB) to overcome/exploit variations in wireless users' capacity per RB. If a user's traffic is leaky bucket constrained, one can determine a fixed service rate that will ensure a desired maximum delay. This permits one to devise a scheduler, the Wireless Guaranteed Rate Service (WGRS), which will ensure a user will see a fixed service rate even though with channels that have stochastic variations.

In fact, any scheduler which allocates at least as much cumulative service as the WGRS scheduler over busy periods is GRS *compliant*, and will thus also meet the user's delay deadlines. This observation suggests the possibility of opportunistically serving a user's data ahead of time when channel rates are good, relative to the GRS scheduler, and/or delay such service when channel rates are poor, as long as the scheduling is GRS compliant. We devise a class of Opportunistic GRS (OGRS) schedulers that take advantage of this relaxation along with knowledge of the statistics of the users' channel variations, to achieve better spectral efficiency while meeting users' strict delay constraints.

By considering *oracle-aided* policies that have access to future channel capacity realizations, we show via extensive simulations that OGRS can be within 10% to 40% of such policies as the delay constraint is relaxed. These gains translate to doubling the eMBB user's throughput even for the weakest user when URLLC and eMBB traffic share resources or an increase of upto 57% in the number of users admitted as long the arrival rates and channel strengths are similar for the newly admitted users.

Finally, we propose a Measurement Based Admission Control (MBAC) strategy, that indirectly accounts for the heterogeneity in traffic, channel, and delay constraints by directly tracking resource usage statistics based on the resource allocation algorithm of our proposed OGRS scheduler. While this approach may be more robust to uncertainty, it may fail from

ISBN 978-3-903176-55-3 © 2023 IFIP

time to time, unlike previously considered MBAC policies, and may have to resort to prioritizing a particular class of users.

II. System Model

We consider discrete time downlink scheduling for a base station serving a set $\mathcal{U}$ of URLLC users with stochastic arrivals and possibly heterogeneous QoS requirements and a set $\mathcal{E}$ of backlogged eMBB users. We denote by $(A_n^u)_{n\in\mathbb{N}}$ the arrival process for user $u \in \mathcal{U}$, where A_n^u is a random variable denoting the number of bits that arrive and are available for service in time slot n with a transmission deadline of $n+d$. In general, it is not possible to ensure delay guarantees to a user without prior knowledge of its traffic statistics or of constraints on its traffic. A common approach for the latter is to establish and enforce (through traffic policing/shaping) apriori constraints on the user's traffic that can be used to design resource allocation mechanisms guaranteed to meet a user's QoS requirements. In Section III and IV of this paper we will assume each user's traffic satisfies dual leaky bucket constraints [9] with parameters $(\rho^u, \sigma^u, \mu^u)$, where σ^u denotes the token bucket size in bits and ρ^u, μ^u denote the peak and mean bit arrival rate per time slot, respectively. The user's cumulative arrival process $A^u(\cdot,\cdot]$ is thus constrained as follows for all $\tau, n \in \mathbb{N}$,

$$A^u(\tau, \tau+n] = \sum_{k=\tau+1}^{\tau+n} A_k^u \leq \min\left[\rho^u n, \sigma^u + \mu^u n\right] . \quad (1)$$

The base station transmit resources are modeled as a sequence of frames/slots each comprising multiple Resource Blocks (RBs) which can be arbitrarily allocated to users on a per time slot basis by the scheduling policy. Each RB denotes a slice of time and frequency block available to the BS for resource allocation. We let the random variable $C_n^u \in \mathbb{R}^+$ denote the channel rate (bits per RB) that can be transmitted to user u if it is allocated a *single* RB on time slot n. A user may be allocated multiple RBs, but we assume a *flat fading* setting where the rate delivered to u is the same across RBs in a given time slot. Further, we assume $(C_n^u)_{n\in\mathbb{N}}$ are independent and identically distributed (i.i.d.) across time slots. A non zero transmission rate C_n can be viewed as a coverage/connectivity requirement for users.

We consider a system model where a scheduling policy, say π, decides the number of RBs be allocated to each user in each time slot. The decision of policy π at time n is assumed to be causal concerning knowledge of the current and past channel rates $(C_\tau^u)_{\tau=0}^n$, arrivals and queue lengths, allowing for opportunistic scheduling, i.e., taking advantage of capacity variations across time. In particular, we let $M_n^{u,\pi} \in \mathbb{R}^+$ denote the number of RBs allocated to user u on slot n by a policy π given the observed history. Such an allocation provides an overall service rate $S_n^{u,\pi}$ (total bits transmitted with potentially multiple RBs allocated) to the user u on time slot n given by,

$$S_n^{u,\pi} = M_n^{u,\pi} C_n^u,$$

and we define the cumulative service over an interval $(\tau, \tau+n]$ as follows,

$$S^{u,\pi}(\tau, \tau+n] = \sum_{k=\tau+1}^{\tau+n} S_k^{u,\pi} . \quad (2)$$

A user's data queue (in bits) is modeled as a First Come First Serve (FCFS) discrete time queue with arrivals A_n^u and service rate $S_n^{u,\pi}$ as shown in Fig. 1. We let $Q_{n+1}^{u,\pi}$ denote the number of bits in the user's queue at the start of slot $n+1$, then

$$Q_{n+1}^{u,\pi} = [Q_n^{u,\pi} - S_n^{u,\pi}]^+ + A_{n+1}^u . \quad (3)$$

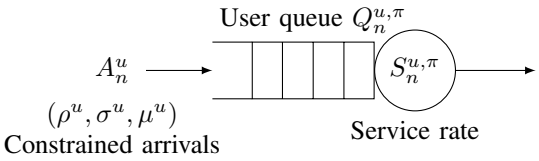

Fig. 1: Leaky bucket constrained arrivals to a discrete time queue with a service rate controlled by scheduling policy π.

III. Guaranteed Rate Scheduling

In this section, we will assume user traffic is leaky bucket constrained, whence assuming a user's data queue served in FCFS order, *it's delay requirement will be met through a sufficiently high fixed service rate per slot*. Note that in practice wireless capacity varies over time, yet we will start by introducing this example where the user's rate is *fixed* and later consider how to address user's channel variability across time. Without loss of generality, we shall henceforth present the analysis for a *single user* with traffic shaping parameters (ρ, σ, μ) and delay requirement d. Referring to the network calculus literature [9], the minimal service rate s required to meet the user's delay constraint d must satisfy,

$$d \leq \frac{[\rho - s]^+}{\rho - \mu} \frac{\sigma}{s} \implies s = \frac{\rho\sigma}{(\rho-\mu)d + \sigma} . \quad (4)$$

where $[x]^+ = \max[x, 0]$. *As long as a user is allocated enough resources to meet the service rate of s, it will meet the packet level delay requirement owing to leaky bucket constrained traffic.* This is easily visualized, see Fig. 2. The red curve is the worst case cumulative arrivals for leaky bucket constrained traffic, the blue line a fixed rate service, and the green interval the worst case delay a bit must wait until service.

Definition 1. ***GRS(s)*** *We let the Guaranteed Rate Scheduler with service rate s, GRS(s), be the scheduling policy that guarantees a user data queue a service rate of at least s per slot whenever it is sufficiently backlogged.*

The above definition matches with that of a strict service curve in Deterministic Network Calculus [9].

Definition 2. ***WGRS(s)*** *A wireless GRS(s) scheduler for a user with time varying channel rate C_n bits per RB allocates a time varying number of RBs M_n to the user such that at time n,*

$$M_n = \min\left[\frac{s}{C_n}, \frac{Q_n}{C_n}\right], \quad (5)$$

ISBN 978-3-903176-55-3 © 2023 IFIP

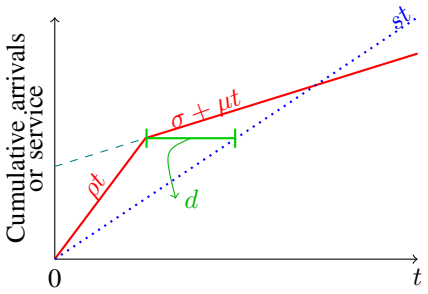

Fig. 2: Leaky bucket constrained flow arrival and service curves for deterministic service rate.

where C_n is the channel rate at time n and Q_n denotes the number of bits in the user's queue, resulting in an overall service rate $S_n = M_n C_n$.

Under these policies, as long as the user's queue is sufficiently backlogged the user will see a service rate s. Since the GRS(s) scheduler satisfies the user's delay constraints for the appropriately selected s, so will the wireless version, although it may require the allocation of a large number of RBs if the user's channel capacity is low. Recall we consider a setting where there is a sufficiently large number of RBs available to users, and they are unlikely to require a lot of resources at the same time. Further note that although the scheduler is designed based on worst case analysis, resources are only be allocated if needed, i.e., only if the user queue is backlogged, hence no resources would be wasted, indeed they will be allocated to other users.

IV. Opportunistic GRS schedulers

Following the setting in Section III, we shall propose a new class of wireless schedulers that we refer to as Opportunistic GRS(s) schedulers that have additional flexibility to exploit temporal variations in users channel capacity, yet are guaranteed to meet the user's delay requirements. To that end, we first formally define a property that ensures the policy will meet the same delay requirements as GRS(s) when users' traffic is leaky bucket constrained.

Definition 3. ***GRS(s) compliant*** *A scheduling policy π is GRS(s) compliant if when subject to the* same *arrivals process for any busy cycle of the GRS(s) scheduler, say $(0, b]$, the cumulative service of π over the interval $(0, \tau]$ for $\tau = 1, 2, \ldots, b$ is greater than or equal to that of the GRS(s) scheduler. It follows that the user queue under GRS(s) compliant policy will empty out whenever that under the GRS(s) scheduler empties.*

Since a GRS(s) compliant policy's cumulative service (departures) is greater than that of the GRS(s) policy on any busy cycle, it is clear that it can only speed up departures and thus reduce delays in FCFS user queues. We note that GRS(s) compliance differs from the traditional service curve definition, see e.g., [9], [10]), in that it is defined via a coupling of π to the GRS(s) policy on busy cycles, and in particular it is not shift invariant, i.e., under π it is possible to have an interval in which queues are backlogged and there are no departures. Fig. 3 exhibits a sample realization. In the figure, the red curve shows the cumulative arrivals to a GRS(s) busy cycle beginning at time 0 – corresponds to the worst case cumulative arrivals associated with a dual leaky bucket. The dotted line represents the cumulative service at a fixed rate s. Meanwhile the blue cumulative service curve corresponds to a policy π. As can be seen, from the start of the busy cycle at time 0 to the end at time b, the cumulative service of policy π exceeds that of the fixed rate service and so GRS(s) compliant.

Fig. 3 also exhibits the perspective underlying Opportunistic GRS scheduling. *The key idea is to exploit temporal channel rate variability to improve spectral efficiency without impacting delay guarantees.* We observe that at times t_1 and t_3 the user's channels are particularly good, and the user has queued data significantly higher than s. The scheduler chooses to exploit these good user channels, by serving much more data at those times than the minimal service rate required by GRS(s) scheduling. In principle, since the user's channel is good at those times, the number of RBs the wireless scheduler would be allocating by doing so would be reduced as compared to the WGRS(s) scheduler introduced in the previous section. Next, we formally introduce a class of Opportunistic GRS(s) scheduling policies.

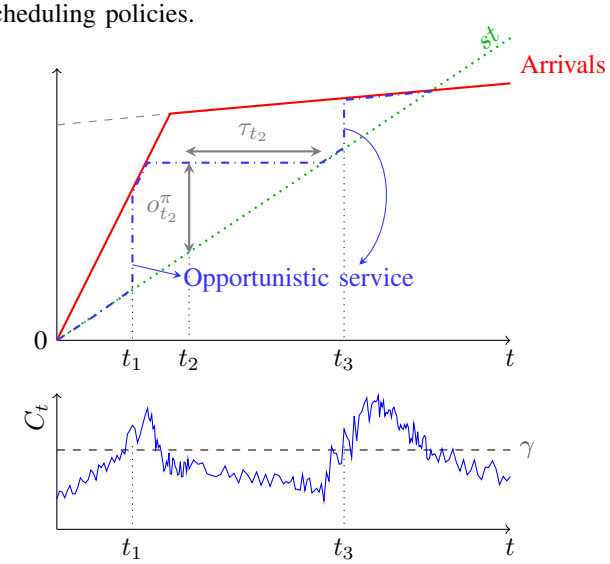

Fig. 3: Temporal channel variations and opportunistic service based on bits in queue. Note that continuous time data rate has been used for ease of understanding the theory.

Definition 4. ***OGRS(s)*** *An Opportunistic GRS(s) scheduling policy π subject to an arrival process $(A_n)_n$ simulates the GRS(s) scheduler and allocates RBs and thus service $S^{\pi}(\cdot, \cdot]$ to the user such that for any GRS(s) busy cycle, say $(0, b]$, we have*

$$A(0, \tau] \geq S^{\pi}(0, \tau] \geq S^{GRS}(0, \tau] \quad \text{for} \quad \tau = 1, 2, \ldots, b,$$

where $A(0, \tau]$ denotes the cumulative arrivals and $S^{GRS}(0, \tau]$ the cumulative service allocated by GRS(s) since the beginning of the busy cycle.

ISBN 978-3-903176-55-3 © 2023 IFIP

By definition, OGRS(s) schedulers are GRS(s) compliant, and thus will satisfy the user's delay requirements if s is chosen appropriately, relative to the arrival processes' leaky bucket parameters. However, such schedulers have the additional freedom to decide when to allocate RBs to the user and in particular, to do so when the channels are particularly good.

Definition 5. ***Threshold-based Opportunistic GRS(s)*** *The basic principle underlying a threshold-based OGRS(s) scheduling policy π is as follows: if on time slot n the channel rate C_n exceeds a threshold γ_n^π, then a sufficient number of RBs are allocated by the scheduler to clear the queue backlog, i.e, $M_n^\pi = Q_n^\pi/C_n$. Otherwise, a minimal number of RBs are allocated so as to ensure the cumulative service allocated keeps up with that of the GRS(s) scheduler over its busy cycles.*

Note, that the threshold γ_n^π can be time/state dependent and controls how the algorithm exploits channel rate fluctuations – this will be explained in the sequel.

Algorithm 1 exhibits the details of the threshold based OGRS(s) scheduler which operates with respect to the cumulative service a virtual GRS scheduler would provide for the same arrival process. To start with, consider a GRS(s) busy cycle that without loss of generality begins at 0. Then one can express the user queue length and service for the GRS(s) at time n as follows,

$$Q_n^{\text{GRS}} = [Q_{n-1}^{\text{GRS}} - s]^+ + A_n,$$
$$S_n^{\text{GRS}} = \min[Q_n^{\text{GRS}}, s].$$

Therefore, the cumulative service provided by GRS(s) over an interval $(0, n]$ can be expressed as,

$$S^{\text{GRS}}(0, n] = S^{\text{GRS}}(0, n-1] + \min[Q_n^{\text{GRS}}, s]\,.$$

Next, we have the OGRS policy which needs a metric that can measure the amount of service provided in excess of the guaranteed rate s per time. slot. Let O_n^π denote the amount of data that has been (opportunistically) sent ahead of time n relative to the GRS(s) scheduler, i.e.,

$$O_n^\pi = S^\pi(0, n-1] - S^{\text{GRS}}(0, n-1].$$

We initialize $O_0^\pi = 0$ at the start of a busy cycle. Note that the duration of a busy cycle of a GRS(s) compliant scheduling policy with leaky bucket constrained arrivals is upper bounded [9] by $b_{\max}(s) = \frac{\sigma}{s-\mu}$, which bounds O_n^π and guarantees it will eventually return to 0.

As mentioned in the policy Definition 5 above, if $C_n > \gamma_n^\pi$ then π serves all the data in the user queue, i.e., $S_n^\pi = Q_n^\pi$. Clearly, the metric O_n^π must be positive if $Q_{n-1}^\pi > s$, because π has served all the traffic that has entered the queue since the start of the busy cycle, while GRS(s) only the bare minimum service it guarantees.

When the channel rate is not so good, i.e., if $C_n \le \gamma_n^\pi$ then OGRS can choose to not schedule any RBs if at least s bits had been transmitted in advance. Specifically, if the amount of excess service at time $n-1$ falls short of s then π only serves the minimum number of bits to ensure it keeps up with the GRS(s) scheduler, i.e.,

$$S_n^\pi = [\min[s, Q_n^\pi] - O_n^\pi]^+\,.$$

Algorithm 1: Guaranteed Rate Scheduling with Opportunism Over Temporal variations

initialize $O_0 = 0, S_0^\pi = 0, S_0^{\text{GRS}} = 0$;
while $n > 0$ **do**
  if $C_n > \gamma_n^\pi$ **then**
    $S_n^\pi = Q_n^\pi$;
  else
    $S_n^\pi = [\min[s, Q_n^\pi] - O_n^\pi]^+$;
  end
  $M_n^\pi = S_n^\pi/C_n$;
  $S_n^{\text{GRS}} = \min[s, Q_n^{\text{GRS}}]$;
  $Q_{n+1}^{\text{GRS}} = Q_n^{\text{GRS}} - S_n^{\text{GRS}} + A_{n+1}$;
  $Q_{n+1}^\pi = Q_n^\pi - S_n^\pi + A_{n+1}$;
  $O_{n+1}^\pi = S^\pi(0, n] - S^{\text{GRS}}(0, n]$;
end

A. OGRS Threshold selection

In this section, we propose various ways to design the thresholds $(\gamma_n^\pi)_n$ driving the behavior of the threshold-based OGRS scheduler. We shall assume that the scheduler has access to $F_C(\cdot)$, the CDF for the users' channel rate variations. We also assume that $F_C^{-1}(\cdot)$ is an appropriately defined inverse CDF. As explained in the sequel, in practice the CDF can be inferred, as in [31] to possibly adapt to changes over time.

1) ST(α): Static threshold: Our first threshold design is a static percentile, i.e., $\gamma_n^\pi = \gamma^\pi$ corresponding to the α-percentile of the channel rate CDF, where $\alpha \in (0, 1)$, so,

$$F_C(\gamma^\pi) = \alpha \implies \gamma^\pi = F_C^{-1}(\alpha). \tag{6}$$

For example, with a choice of $\alpha = 0.8$ the OGRS(s) triggers an opportunistic scheduling of the user's queued data only if the current channel rate has exceeded the 80^{th} percentile, i.e., $c_n > \gamma^\pi$. Note that the choice percentile α is a design parameter that can in principle be optimized to minimize the mean resources (RBs) allocated by the associated OGRS(s) scheduler.

2) DTP(δ): Dynamic threshold based on probability: Next we consider thresholds based on a dynamic percentile of the channel rate CDF $F_C(\cdot)$. Recall that $O_n^\pi = o_n^\pi$ denotes the amount of data that our OGRS(s) policy has delivered *ahead* of time as compared to GRS(s) at time n. Given that the GRS(s) must serve at least s bits per slot, an OGRS(s) policy could in principle wait for $\tau_n = \left\lfloor \frac{o_n^\pi}{s} \right\rfloor$ time slots before the GRS(s) scheduler catches up and is forced to schedule at time $\tau_n + 1$. Ideally the data should be scheduled on slot n if the current rate realization c_n is better than that to be observed in the next $\tau_n + 1$ time slots with high probability, i.e.,

$$\mathbb{P}\left(c_n > \max_{i=1,\ldots,\tau_n+1} C_{n+i}\right) \ge \delta. \tag{7}$$

The following lemma translates the above requirement to a threshold on c_n. Note that δ is a design parameter that needs to be carefully chosen so as to minimize the number of RBs required.

Lemma 1. *Let $(C_n)_n$ be i.i.d random variables with the same marginal distribution $F_C(\cdot)$ and appropriately defined inverse $F_C^{-1}(\cdot)$. If c_n exceeds the threshold $F_C^{-1}\left(\delta^{\frac{1}{\tau_n+1}}\right)$ then (7) is satisfied.*

Proof. Since $F_C(\cdot)$ is non-decreasing and recalling that $F_C(C_i) \sim U_i$ are i.i.d. Uniform$[0,1]$ we have that

$$c_n > \max_{i=1,\ldots,\tau_n+1} C_{n+i} \iff F_C(c_n) > \max_{i=1,\ldots,\tau_n+1} U_i, \quad (8)$$

and it follows (7) can be rewritten

$$\mathbb{P}\left(F_C(c_n) > \max_{i=1,\ldots,\tau_n+1} U_i\right) \implies F_C(c_n) > \delta^{\frac{1}{\tau_n+1}}, \quad (9)$$

giving the desired threshold $c_n > F_C^{-1}\left(\delta^{\frac{1}{\tau_n+1}}\right)$. □

3) DTE: Dynamic threshold based on expectation: A user with current channel rate c_n might choose not to schedule transmissions on the current slot in the hope of seeing a better channel in the next τ_n slots. The previous threshold design was based on the inequality in (7) being satisfied with high probability. Alternatively, the current channel rate c_n might be considered good if one can ensure the inequality holds on average. Taking the expectation of the inequality on the right hand side of and computing the expectation of the max of uniform random variables in (8) gives,

$$F_C(c_n) > \mathbb{E}\left[\max_{i=1,\ldots,\tau_n+1} U_i\right] = 1 - \frac{1}{\tau_n+2}.$$

Under this rough approximation an associated threshold on c_n depends on τ_n which can be set to,

$$\gamma_n^{\pi} = F_C^{-1}\left(1 - \frac{1}{\tau_n+2}\right). \quad (10)$$

This captures the key insight that with a larger number of slots τ_n, an OGRS(s) scheduler can choose to wait until the channel rate exceeds the $1 - 1/(\tau_n+2)$ percentile. Furthermore, this threshold selection mechanism does not have any design parameter which makes it easier to implement in practice.

V. Simulation Results

We consider a BS serving a set of URLLC and eMBB users with each user's channel rate C_n per RB determined by the corresponding received Signal to Noise Ratio (SNR). The received SNR was modelled using the 3GPP Urban-Micro path loss model [1], with Rayleigh distributed small scale fading. We assume bounded channel realizations, where the SNR lies between -6.934 dB $\leq$ SNR $\leq$ 20 dB. For simplicity, we shall use Shannon capacity $B\log_2(1+\text{SNR}_n)$ to calculate the rate per RB C_n, where each RB is a time frequency slice of duration 1ms with bandwidth $B = 10$ KHz. We initially use Shannon capacity $B\log_2(1+\text{SNR}_n)$ to calculate the rate per RB C_n, where each RB is a time frequency slice of duration 1ms with bandwidth $B = 10$ KHz. Later on, we demonstrate our performance using realistic 3GPP modulation and coding scheme [1] taking one of 15 quantized values in -6.934 dB $\leq$ SNR $\leq$ 20 dB. Finally, to determine the channel quality thresholds, we need the CDF $F_C(\cdot)$ for the channel rate per RB on a given slot, for each user. We used the last 100 channel SNR realizations to determine the empirical CDF of the user's SNR at any given instant. Note that all plots in this section were generated over 10^6 slots, resulting in a ± 0.1 error for the estimated mean number of allocated RBs per slot $\bar{M}^{\pi}$ with 99% confidence interval. Additionally, we will use the WGRS policy as a baseline and also the multicarrier version of MLWDF [8] policy as a benchmark for evaluating the performance of our proposed algorithm.

A. Performance of opportunistic scheduling policies

1) URLLC spectral efficiency: For this subsection we consider three different URLLC users that are at distances 300, 500 and 700 meters and which we refer to as strong, medium and weak users, respectively. Also, we assume each URLLC user has stochastic arrivals with packet size of 1024 bits that arrive each time slot shaped by the leaky bucket with parameters (in packets per time slot) $\sigma = 50, \rho = 10, \mu = 5$. The guaranteed packet rate s per time slot is then determined using (4) for a common delay deadline of $d = 4$ ms for all the users. The efficiency of our proposed scheudling policy is measured in terms of the average number of RBs required to serve the user, subject to its delay deadline. Henceforth, we shall refer to the proposed OGRS scheduling policy by the threshold method that we employ to determine the channel rate quality.

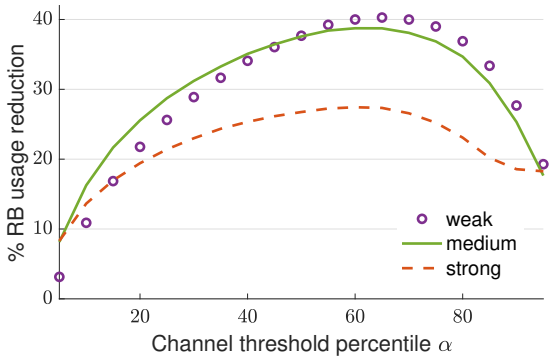

Fig. 4: Percentage reduction in RB usage for ST(α) with respect to (w.r.t) that of WGRS.

Fig. 4 shows the percentage reduction in the average number of RBs required to serve all three types of users with ST(α) relative to that needed by WGRS under the same arrival and channel rate processes. It is interesting to note that the weak user sees the best gain, which can be attributed to the higher range of channel strength variations that a typical weak user would observe. Higher temporal variability should lead to higher opportunistic gains for such users. Also, it is clear that

ISBN 978-3-903176-55-3 © 2023 IFIP

the percentile α that maximizes the spectral efficiency gain depends on the arrival pattern and user channel strength.

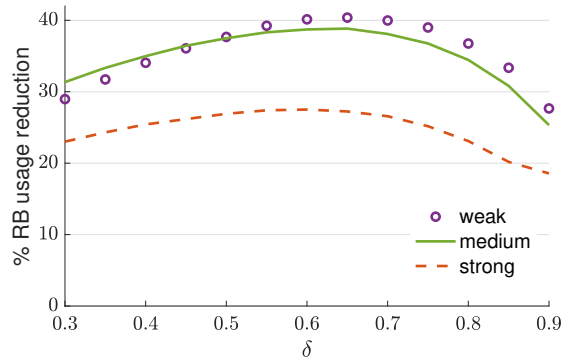

Fig. 5: Resource reduction percent for DTP(δ) w.r.t. WGRS.

Finally, we also evaluated the efficiency gains of DTE vs WGRS(s) and the results were as follows: the percentage reduction in the number of resources required for each type of strong, medium, and weak, users were $26.79\%, 37.47\%, 37.67\%$ for the strong, medium and weak user, respectively. The DTE policy achieves at most 5% loss in efficiency as compared to DTP(δ) policy, thus providing a reasonable approach as it requires no parameter fine tuning.

2) Improvement in eMBB throughput: In this subsection, we demonstrate the throughput improvement for eMBB users where the BS supports multiple heterogeneous users, 3 URLLC and 5 eMBB users. We consider ON-OFF bursty arrivals for URLLC users, where packets arrive at a peak rate ρ during the ON period. The ON, OFF cycles are of duration $\frac{\sigma}{\rho-\mu}, \frac{\sigma}{\rho}$, respectively. Distance from the BS and leaky bucket parameters for the 3 URLLC users are tabulated below:

TABLE I: Leaky bucket parameters for multiple users.

User	distance (m)	Delay(ms)	ρ	μ	σ
1	300	5	10	5	50
2	500	3	20	10	50
3	700	7	10	5	50

The eMBB users are located at distances $250, 560, 650, 720, 800$ meters from the BS. Note that we assume that eMBB users are infinitely backlogged and do not have any stringent deadlines. Furthermore, we use proportionally fair scheduling to select one eMBB user from the set of eMBB users at each time slot that gets allocated all the leftover RBs.

A total of 6000 RBs are available to all the users connected to the BS and the URLLC users are allocated resources with priority. After allocating resources to all active URLLC users, the leftover RBs are used to serve eMBB users. The throughput performance of the oracle-aided policy is also included to provide a *bound* on the best feasible spectral efficiency for URLLC users, which translates to higher throughput for eMBB users.

Fig. 6 showcases the throughput gains for eMBB users for the various algorithms. Clearly, when compared to the baseline and benchmark scheduling policies, OGRS policy is indeed closer to the throughput gain bound set by the oracle-aided scheduling policy with access to future channel rates.

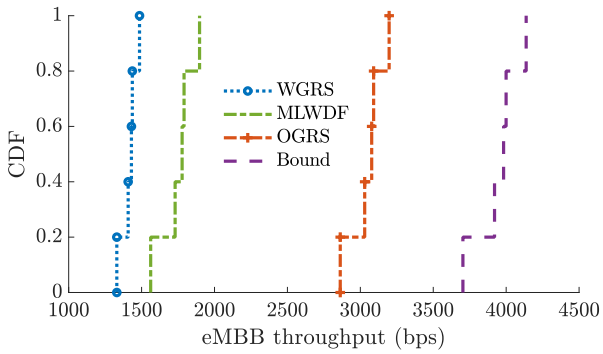

Fig. 6: Long term throughput distribution for eMBB users.

3) Sensitivity to history: We observed that for as little as 10 samples of the past channels, we were within a 5% of the spectral efficiency when using 100 samples for estimating the CDF. Clearly, a short history such as ten past samples is sufficient to track the wireless channel variations without significant loss in the efficiency of our scheduling algorithms.

B. Admission Control

We consider a set of 100 users with ON-OFF bursty traffic and leaky bucket constrained arrivals. The delay deadline and user location (distance from the BS) are drawn uniformly random from the sample spaces $\{200, 250, \ldots, 800\}$ and $\{2, 3, \ldots, 10\}$, respectively. The arrivals and channel variations are generated over 10^6 time slots to simulate the number of users that can be admitted for various system capacities $\overline{m}$, the total number of RBs available in the system. For the same set of users, we also use the Gaussian approximation for the aggregate resource requirement X_u to determine the number of admitted users, shown as dashed lines in Fig. 7. If Y is the random variable that denotes the resource requirement of a new user, then the probability that the total resource requirement $X_u + Y$ will exceed $\overline{m}$ is approximated using the following inequality,

$$\mathbb{P}\left(X_u + Y > \overline{m}\right) \leq \exp\left(-\frac{(\overline{m}-\mu)^2}{2\sigma^2}\right), \qquad (11)$$

where $\mu = \mu_u + \hat{\mu}$ and $\sigma^2 = \sigma_u^2 + \hat{\sigma}^2$. Note that the inequality in (11) provides a computationally reasonable expression that can be used to decide if the new user can be admitted without exceeding the reliability requirement δ.

It can be seen that the Gaussian approximation provides a conservative estimate of the number of users that can be admitted into the system for both the OGRS and WGRS

ISBN 978-3-903176-55-3 © 2023 IFIP

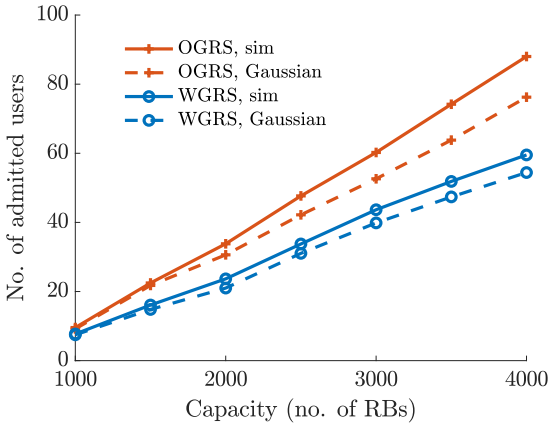

Fig. 7: Admission control using Gaussian approximation on the total RBs required for all admitted URLLC users.

scheduling policies. Note that OGRS scheduling policy is able to admit more users when compared to the WGRS policy, which is interesting given that WGRS is more deterministic in resource provisioning, whereas the others are more bursty, as a function of the channel quality and arrivals.

VI. Conclusion

We have proposed a measurement based opportunistic wireless scheduler, which can meet heterogeneous users' hard delay deadlines while being spectrally efficient, i.e., minimizing the resources required, thus permitting the system to achieve additional throughput for other traffic sharing the network resources. The underlying design principle for OGRS policies is to ensure that the wireless scheduler meets or exceeds the service that a guaranteed rate scheduler with rate s would assign. Thus by design, OGRS policies can also be used to efficiently deliver a Guaranteed Bit Rate (GBR) service. Our proposed policy uses dynamic opportunistic thresholds to leverage the knowledge of the user's marginal channel quality rate distribution, which in practice would be measured and/or tracked based on a limited number, say 10, of the previous channel realizations.

Acknowledgement

We would like to thank our peers at Samsung, an affiliate of the 6G@UT center within the Wireless Networking and Communications Group at The University of Texas at Austin.

References

[1] 3GPP, "5G Study on Channel Models for frequencies from 5 to 100 GHz Release 14," Tech. Report (TR) 38.901, (3GPP), 04 2018. Version 14.3.0.

[2] 5G Americas, "New Services & Applications with 5G Ultra-Reliable Low Latency Communications," White Paper, 2018.

[3] A. Jalali *et al.*, "Data Throughput of CDMA-HDR a High Efficiency-High Data Rate Personal Communication Wireless System," in *IEEE 51st Veh. Tech. Conf. Proc.*, vol. 3, pp. 1854–1858, 2000.

[4] A. Eryilmaz *et al.*, "Discounted-Rate Utility Maximization (DRUM): A Framework for Delay-Sensitive Fair Resource Allocation," in *15th Int. Symp. on Modeling and Opt. in Mobile, Ad Hoc, and Wireless Netw. (WiOpt)*, pp. 1–8, 2017.

[5] L. Tassiulas *et al.*, "Stability Properties of Constrained Queueing Systems and Scheduling Policies for Maximum Throughput in Multihop Radio Networks," in *29th IEEE Conf. on Decision Control*, pp. 2130–2132, 1990.

[6] S. Shakkottai *et al.*, "Scheduling for Multiple Flows Sharing a Time-Varying Channel: The Exponential Rule," *Translations of the American Mathematical Society-Series 2*, vol. 207, pp. 185–202, 2002.

[7] B. Sadiq *et al.*, "Delay-Optimal Opportunistic Scheduling and Approximations: The Log Rule," *IEEE/ACM Trans. on Netw.*, vol. 19, no. 2, pp. 405–418, 2011.

[8] M. Andrews *et al.*, "Providing Quality of Service Over a Shared Wireless Link," *IEEE Commun. Mag.*, vol. 39, no. 2, pp. 150–154, 2001.

[9] J.-Y. Le Boudec *et al.*, *Network Calculus: A Theory of Deterministic Queuing Systems for the Internet*. Lec. Notes in Comp. Sci., Springer Berlin Heidelberg, 2003.

[10] C.-S. Chang, *Performance Guarantees in Communication Networks*. Berlin, Heidelberg: Springer-Verlag, 2000.

[11] L. Le *et al.*, "Service Differentiation in Multirate Wireless Networks with Weighted Round-Robin Scheduling and ARQ-based Error Control," *IEEE Trans. on Commun.*, vol. 54, no. 2, pp. 208–215, 2006.

[12] P. Lin *et al.*, "CS-WFQ: a Wireless Fair Scheduling Algorithm for Error-Prone Wireless Channels," in *Proc. 9th Int. Conf. on Comp. Commun. and Netw.*, pp. 276–281, 2000.

[13] S. Patil *et al.*, "Managing Resources and Quality of Service in Heterogeneous Wireless Systems Exploiting Opportunism," *IEEE/ACM Trans. on Netw.*, vol. 15, no. 5, pp. 1046–1058, 2007.

[14] J. J. Jaramillo and R. Srikant, "Optimal Scheduling for Fair Resource Allocation in Ad hoc Networks with Elastic and Inelastic Traffic," in *2010 Proceedings IEEE INFOCOM*, pp. 1–9, IEEE, 2010.

[15] C. Tsanikidis *et al.*, "On the Power of Randomization for Scheduling Real-Time Traffic in Wireless Networks," in *IEEE INFOCOM 2020 - IEEE Conf. on Comp. Commun.*, pp. 59–68, 2020.

[16] B. Li *et al.*, "Wireless Scheduling Design for Optimizing Both Service Regularity and Mean Delay in Heavy-Traffic Regimes," *IEEE/ACM Trans. on Netw.*, vol. 24, no. 3, pp. 1867–1880, 2016.

[17] A. Destounis *et al.*, "Scheduling URLLC Users with Reliable Latency Guarantees," in *16th Int. Symp. on Modeling and Opt. in Mobile, Ad Hoc, and Wireless Netw. (WiOpt)*, pp. 1–8, 2018.

[18] P. Ameigeiras *et al.*, "3GPP QoS-based scheduling framework for LTE," *EURASIP Journal on Wireless Commun. and Netw.*, pp. 1–14, 2016.

[19] A. Anand *et al.*, "Joint Scheduling of URLLC and eMBB Traffic in 5G Wireless Networks," in *IEEE INFOCOM 2018 - IEEE Conf. on Comp. Commun.*, pp. 1970–1978, 2018.

[20] A. A. Esswie and K. I. Pedersen, "Opportunistic Spatial Preemptive Scheduling for URLLC and eMBB Coexistence in Multi-User 5G Networks," *IEEE Access*, vol. 6, pp. 38451–38463, 2018.

[21] A. Karimi *et al.*, "Efficient Low Complexity Packet Scheduling Algorithm for Mixed URLLC and eMBB Traffic in 5G," in *IEEE 89th Veh. Tech. Conf. (VTC2019-Spring)*, pp. 1–6, 2019.

[22] H. Yin *et al.*, "Multiplexing URLLC Traffic Within eMBB Services in 5G NR: Fair Scheduling," *IEEE Trans. on Commun.*, vol. 69, no. 2, pp. 1080–1093, 2021.

[23] M. Alsenwi *et al.*, "eMBB-URLLC Resource Slicing: A Risk-Sensitive Approach," *IEEE Commun. Lett.*, vol. 23, no. 4, pp. 740–743, 2019.

[24] A. Manzoor *et al.*, "Contract-Based Scheduling of URLLC Packets in Incumbent EMBB Traffic," *IEEE Access*, vol. 8, 2020.

[25] D. I. Shuman *et al.*, *Opportunistic Scheduling with Deadline Constraints in Wireless Networks*, pp. 127–155. Springer New York, 2011.

[26] H. Wu *et al.*, "Application-Level Scheduling With Probabilistic Deadline Constraints," *IEEE/ACM Trans. on Netw.*, vol. 24, no. 3, 2016.

[27] R. J. Gibbens *et al.*, "Measurement-based Connection Admission Control," in *15th Int. Teletraffic Congress*, vol. 2, pp. 879–888, 1997.

[28] M. Grossglauser *et al.*, "A Framework for Robust Measurement-based Admission Control," *IEEE/ACM Trans. on Netw.*, vol. 7, no. 3, pp. 293–309, 1999.

[29] D. Tse *et al.*, "Measurement-based call admission control: Analysis and simulation," in *Proceedings of INFOCOM'97*, vol. 3, pp. 981–989, IEEE, 1997.

[30] L. Breslau *et al.*, "Comments on the Performance of Measurement-based Admission Control Algorithms," in *Proc. IEEE INFOCOM 2000. Conf. on Comp. Commun.*, vol. 3, pp. 1233–1242, 2000.

[31] S. Patil *et al.*, "Measurement-Based Opportunistic Scheduling for Heterogeneous Wireless Systems," *IEEE Trans. on Commun.*, vol. 57, p. 2745–2753, sep 2009.