Delay and Jitter Constrained Wireless Scheduling with Near-Optimal Spectral Efficiency

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 devices/applications in terms of their traffic characteristics and service requirements. Particularly challenging is the need to deliver traffic subject to delay and reliability constraints in a spectrally efficient manner. We propose a new measurement-based Opportunistic Guaranteed Deadline Scheduler (OGDS) that meets strict delay deadlines on users' packets. This is achieved by scheduling packet transmissions when the current channel rate is better than that expected in the time window before packet deadlines expire. In order to meet such requirements one must have a complementary admission control policy. We exhibit a simple, once again measurement based policy, that indirectly accounts for heterogeneity in traffic, channel and delay constraints by monitoring statistics of OGDS's resource usage. We show via extensive synthetic and trace driven simulations that OGDS requires at most $10-25\%$ more resources compared to an optimal *offline* scheduling policy with complete knowledge of future channel rates, and performs much better than standard baselines including the state-of-the-art MLWDF scheduler. Finally, we propose a modification to OGDS that enables one to control the jitter at a possible loss in spectral efficiency.**

***Index Terms*—Wireless resource allocation, Delay deadline, Opportunistic scheduling, Low Latency, QoS constraints**

I. Introduction

Each generation of wireless technology has pushed downlink/uplink data rates higher to support ever increasing numbers of devices and applications with higher data requirements. In particular, 5G/6G standards have sought to not only deliver high data rates but also manage users' heterogeneous Quality of Service (QoS) in terms of delay, reliability, jitter, spectral efficiency and throughput. There is a substantial literature in wireless (and wireline) scheduling that provides different tools to address the above challenge, yet, as discussed below, it still falls short in many respects. Below, we briefly highlight some of that literature and the associated shortcomings.

A. Related work

Many works have focused on a setting where users' data queues are *fully backlogged.* When this is the case, one can consider devising schedulers that maximize the sum of the users' utility of their allocated long term rate [1]. For example, Proportionally Fair (PF) wireless scheduling emerges when users have log utility functions, see e.g., [2], and results in a scheduler that realizes a good tradeoff between *opportunistically* scheduling users which have good channels versus achieving a *fair* long term allocation amongst the users. Variations on these ideas have been proposed where the users' utility is a function of the short term throughput, see e.g., [3]. This leads to a more responsive allocation avoiding short term neglect of any user. In practice, PF, and other utility maximizing schedulers, provide a simple and effective strategy for best effort or enhanced Mobile Broadband (eMBB) traffic with no strict delay requirements. Still, questions remain as to what happens when user queues are not fully backlogged or how to choose the fairness criterion, i.e., utility functions when there are delay constraints that require high reliability.

In settings where users' queues are not fully backlogged, researchers have focused on devising queue and channel dependent wireless schedulers which are *throughput optimal*, i.e., ensure user queues' stability whenever feasible. These schedulers also address performance objectives, such as Max-Weight [4], which is delay optimal in the idealized symmetric case, Exponential rule [5] which attempts to minimize the max user queue, and Log rule [6] which attempts to minimize the mean delay. Such schedulers have been adapted to more practical settings, such as the Modified Largest Weighted Delay First (MLWDF) [7] which schedules users based on head-of-line packet delays, current channels, and other hyperparameters reflecting user QoS/allocation objectives. In practice, such schedulers do meet delay constraints (with high probability) if sufficient resources have been provisioned, yet it is difficult to verify when this is true, and as such provide a graceful degradation across users when this is not the case.

Another class of wireless schedulers was born from modifying/adapting ideas from wireline scheduling (e.g., traffic shaping and network calculus [8], [9]) to meet QoS requirements under wireless channel variations. For instance, weighted round robin [10] or weighted fair queueing [11] employ user weights drawn from heuristics or tokens [12] based on service deficit [13] 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, e.g., best effort users sensitive to throughput, or traffic that is sensitive to packet delays. In practice, wireless systems need to be shared by heterogeneous user classes.

While many schedulers in the existing literature address delay constraints for real time traffic, spectral efficiency is often neglected, leading to lesser resource availability for non real time traffic. In this paper, we place such interplay front and center, with a focus on not only developing a scheduler

that meets delay constraints but one that does so in a spectrally efficient manner.

No practical wireless scheduling policy is complete without a complementary strategy for admission control and/or traffic shaping. 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 literature that propose Measurement Based Admission Control (MBAC) [14]–[16], it has to be noted that none of those accurately meet the packet loss targets [17] under finite buffer sizes, which is sufficiently similar to delay violation probability.

B. Our contributions

In this paper, we propose and evaluate a new wireless scheduler, Opportunistic Guaranteed Deadline Scheduler (OGDS), that meets packet deadline and/or jitter constraints. We develop measurement based adaptive thresholds on channel rate that drive scheduling decisions in a manner that is sensitive to users' heterogeneous channel, traffic, and packet delay deadlines. *The key idea is to track recent channel rate variations to predict if the current channel state gives the best possible rate to transmit a packet before its deadline expires.* When coupled with an effective admission control policy, our scheduling policy provides excellent spectral efficiency as compared to the spectral efficiency of any optimal offline delay constrained scheduler. Moreover, the performance and flexibility of our algorithm are significantly better than state-of-the-art scheduling policies such as MLWDF [7]. We further propose a modification of OGDS that would allow service providers to control both packet delay and jitter at a possible cost in spectral efficiency.

We demonstrate the performance and robustness of our scheduler on key performance metrics through extensive synthetic and trace driven simulations, i.e., based on 3GPP channel models as well as real world wireless channel time series [18]. Our simulations show that under OGDS, for a fixed set of delay constrained users, one can achieve significantly improved spectral efficiency than MLWDF, in turn allowing one to achieve higher throughput for best effort traffic sharing the same resources. Finally, we propose a measurement based admission control strategy, which captures the traffic dynamics, channel variations, and the scheduler's resource allocation strategy. This circumvents the need for a model for QoS prediction, but not unlike previously considered MBAC or other admission control policies may occasionally fail to meet requirements and have to resort to prioritizing a graceful degradation of users' QoS.

II. System Model

We consider discrete time downlink scheduling for a base station serving a variety of users with either real time or best effort traffic. We denote by set $\mathcal{U}$ the delay constrained users with stochastic arrivals and possibly heterogeneous QoS requirements. The base station also serves a set $\mathcal{E}$ of infinitely backlogged best effort traffic 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 $n + d^u$, where d^u is the delay constraint for user u. 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 this paper, we devise a scheduler that meets packet delay constraints without directly relying on traffic shaping constraints, but assuming admission control is in place.

The 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 non zero transmission rate C_n can be viewed as a coverage/connectivity requirement for users, which is required to provide high levels of reliability to delay constrained traffic. 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. We will consider more general settings such as non identically distributed or correlated channels in the simulations section.

A scheduling policy π, decides the number of RBs to be allocated to each user in each time slot. For ease of exposition, we will assume that there are enough RBs to provision service to all users in the system, and in the sequel, we will introduce admission control to limit the number of users as needed. The decision of policy π at time n is assumed to be causal with respect to 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,

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

The cumulative service over an interval $(\tau, \tau+n]$ is as follows,

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

We let A_n^u, $Q_n^{u,\pi}$, and $S_n^{u,\pi}$ denote the user's arrivals, queue length, and service rate, respectively, at time n. A user's data

queue (in bits) is modeled as a First Come First Serve (FCFS) discrete time queue. Then the number of bits in the user's queue at the start of slot $n+1$, then

$$Q^{u,\pi}_{n+1} = [Q^{u,\pi}_n - S^{u,\pi}_n]^+ + A^u_{n+1}\,. \tag{2}$$

III. Opportunistic Guaranteed Deadline Scheduling

In this section, for the sake of brevity, we will drop the user index (marked by superscript u) as we consider per user scheduler. The user index will be reintroduced in the sequel when we introduce admission control. Suppose we start with an empty user queue at $t=0$, then the arrival process $A(0,\tau]$ delayed by d would be such that $A(-d,\tau-d] = A(0,\tau-d]$. Fig. 1 depicts the cumulative arrivals $A(0,\tau]$ in blue and the corresponding delayed version in red $A(0,\tau-d]$. The delayed arrivals curve represents the worst case cumulative service that a server could provide without violating the delay constraint on each packet. Any cumulative service curve that lies within the arrivals and worst case departures curve will be delay compliant.

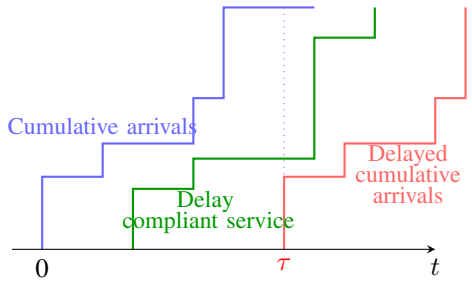


Fig. 1: Delay compliant cumulative service along with worst case delayed service curve.

Definition 1. ***GDS(d)*** *We let the Guaranteed Deadline Scheduler with parameter d, GDS(d), be a scheduling policy that guarantees each bit in the user data queue be serviced within a delay of d since its arrival. Clearly,*

$$A(0,\tau] \geq S^{GDS}(0,\tau] \geq A(-d,\tau-d].$$

Definition 2. ***Opportunistic GDS(d).*** *A threshold based OGDS(d) scheduling policy π is as follows: whenever the channel rate C_n exceeds a threshold γ^π_n, a sufficient number of RBs are allocated by the scheduler to completely clear the queue backlog, i.e, $M^\pi_n = Q^\pi_n/C_n$. Otherwise, a minimal number of RBs are allocated so as to ensure that the cumulative service of π at slot n exceeds or matches that of the d delayed cumulative arrival curve.*

At each time slot, the number of slots τ_n over which there is flexibility to pick when to serve the data in the user queue depends on the residual time until the earliest deadline. Note that while any data whose deadline is due to expire at a given time slot will need to be allocated resources if the current channel rate is expected to be better than those in the next τ_n+1 window (exceeds the threshold), the entire queue backlog is cleared.

Algorithm 1 details the steps involved in OGDS(d) scheduling. When the channel rate is above a certain threshold $C_n > \gamma^\pi_n$, the OGDS policy π serves all data in the user queue,

$$S^\pi_n = Q^\pi_n.$$

Otherwise, the scheduler π allocates only the minimum number of RBs required to meet the worst case delayed service curve, i.e.,

$$S^\pi_n = [A(0,n-d] - S^\pi(0,n-1]]^+\,.$$

Specifically, if the cumulative service provided by π until time $n-1$ is greater than that of the worst case delayed cumulative service at time n, then policy π can completely refrain from allocating any resources at time n if the channel rate is below the threshold.

OGDS Threshold selection: Define τ_n as the slack available to the scheduler before it is forced to schedule data to maintain delay guarantees, i.e.,

$$\tau_n = \min\left[k : k \geq n, A(0,k-d] \geq S^\pi(0,k]\right]. \tag{3}$$

At the time n, the OGDS scheduler has a slack of τ_n time slots before it is forced to start servicing the user queue. Therefore, the current channel rate realization c_n is considered good for opportunistic scheduling if,

$$\begin{aligned} & c_n > \max_{i=1,\ldots,\tau_n+1} C_{n+i}, \\ \iff \quad & F_C(c_n) \overset{(a)}{>} F_C\left(\max_{i=1,\ldots,\tau_n+1} C_{n+i}\right), \\ \iff \quad & F_C(c_n) \overset{(b)}{>} \max_{i=1,\ldots,\tau_n+1} F_C\left(C_{n+i}\right), \\ \iff \quad & F_C(c_n) \overset{(c)}{>} \max_{i=1,\ldots,\tau_n+1} U_i. \end{aligned} \tag{4}$$

Step (a) follows from the monotonicity of the cumulative distribution function (CDF) $F_C(\cdot)$ of the wireless channel rate

Algorithm 1: Guaranteed Deadline Scheduling with opportunism over temporal variations.

```
initialize S^π_0 = 0;
while n > 0 do
    τ_n = min[k : k ≥ n, A(0,k−d] ≥ S^π(0,k]];
    if C_n > γ^π_n then
        S^π_n = Q^π_n ;
    else
        S^π_n = [A(0,n−d] − S^π(0,n−1]]^+ ;
    end
    M^π_n = S^π_n / C_n;
    Q^π_{n+1} = Q^π_n − S^π_n + A_{n+1} ;
end
```

and step (b) follows from the commutative property of the max function with CDF $F_C(\cdot)$. Step (c) follows from the fact [19] that $F_C(C_{n+i}) \sim U_i$ are i.i.d. Uniform$[0,1]$. Since the expectation of the maximum of τ_n+1 independent uniformly distributed random variables is $1-\frac{1}{\tau_n+2}$, a reasonable threshold to determine if c_n is better than the next τ_n channel rate realizations is given by,

$$\gamma_n^\pi = F_C^{-1}\left(1-1/(\tau_n+2)\right). \tag{5}$$

In the discussion above, we have assumed that the user's channel rate CDF is available. Typically, the serving BS tracks the user's Channel State Information (CSI) for adaptive modulation and coding, therefore, it is reasonable to assume that we can empirically estimate the channel rate CDF using CSI [20]. Also note that when the channel rates are discrete, we could use linear interpolation to invert the empirical CDF and compute the percentiles for the channel rate threshold.

A. Modified OGDS

While most scheduling policies for delay constrained traffic focus on improving key performance metrics such as energy efficiency [21], reliability [22] and delay, jitter is often neglected. It is a particularly important metric when transmitting periodic updates to networked real-time control and/or interactive AR/VR gaming applications. Disparate transmission delays across users can be undesirable/intolerable, especially in scenarios that need synchronization of updates across all users. There are multiple ways in which one could measure the variability of transmission delay. In this work, we define jitter in terms of the standard deviation of delay for data transmissions that are periodic in nature.

We propose an elementary modification to the OGDS algorithm that provides a way to trade off between spectral efficiency, delay and jitter, by carefully selecting a transmission window over which resources are allocated to the user. One could either wait a predetermined number of transmit instants, say ζ, or artificially advance the targeted delay deadline to $d-\zeta$ for reducing packet jitter. A shorter window for transmission reduces the number of opportunities available for a user to be efficient, nevertheless, it reduces the variability in delay. Specifically, in Algorithm 1, step 10, the user queue update equation could be modified as follows,

$$Q_{n+1}^\pi = Q_n^\pi - S_n^\pi + A_{n+1+\zeta}\,, \tag{6}$$

where S_n^π denotes the service provided at time n, and $A_{n+1+\zeta}$ stands for the arrivals at time $n+1+\zeta$. In the sequel, we will refer to the parameter ζ as the jitter control parameter and demonstrate how the modified OGDS policy performs in terms of spectral efficiency and jitter.

B. Lower Bound on Spectral Efficiency

In this subsection, we develop a lower bound on the *minimum number of resource blocks* required by any wireless scheduler meeting the delay deadlines. The lower bound is based on considering an *offline* policy with complete knowledge of the future channel realizations and thus not achievable in an online setting, yet a good benchmark.

Consider a user with an arrival process $(A_n)_n$ and a time varying channel rate $(C_n)_n$ per resource block, whose traffic is subject to a delay constraint of at most d slots.

Theorem 1. *For any scheduling policy π meeting the delay constraint, let N_n^π denote the (possibly fractional) number of resource blocks used to serve the arrivals A_n, these RBs may be allocated at the earliest on slot n, but no later than the deadline $n+d$. Similarly, we let M_n^π denote the total number of RBs allocated on slot n. It then follows that,*

$$N_n^\pi \geq A_n \min_{0\leq j\leq d}\left[\frac{1}{C_{n+j}}\right] \ \textit{a.s.}\,. \tag{7}$$

Furthermore, if the arrivals and channel rate processes are stationary and independent of each other and the policy π is such that,

$$\lim_{n\to\infty}\frac{1}{n}\sum_{\tau=1}^{n} N_\tau^\pi = \bar{N}^\pi,$$

then the time average of $(M_n^\pi)_n$ also converges to a limit $\bar{M}^\pi$, which satisfies

$$\bar{M}^\pi = \bar{N}^\pi \geq \mathbb{E}[A_1]\,\mathbb{E}\left[\frac{1}{\max\limits_{0\leq j\leq d} C_{1+j}}\right]. \tag{8}$$

Proof. For any policy π satisfying the delay constraint d, it must be the case that the A_n bits arriving to the user queue at time n are served within the next d slots. Thus, in particular, if we let $S_{n+j}^{\pi,n}$ denote the number of bits of A_n that are served on slot $n+j$, we have that the possibly fractional number of resource blocks required must satisfy,

$$\begin{aligned} N_n^\pi &= \sum_{j=0}^{d}\frac{S_{n+j}^\pi}{C_{n+j}},\ \text{where}\ \sum_{j=0}^{d} S_{n+j}^\pi = A_n, \\ &\geq A_n \min_{0\leq j\leq d}\left[\frac{1}{C_{n+j}}\right] \quad \text{a.s.} \end{aligned} \tag{9}$$

It is easy to establish the following inequalities,

$$\frac{1}{n}\sum_{\tau=1}^{n-d} N_\tau^\pi \leq \frac{1}{n}\sum_{\tau=1}^{n} M_\tau^\pi \leq \frac{1}{n}\sum_{\tau=1}^{n} N_\tau^\pi, \tag{10}$$

because on the one hand, the total RBs allocated across the first n time slots is lower bounded by the total number that was allocated to serve the traffic that arrived within $(0, n-d]$; indeed given the delay constraint, all such traffic should be served prior to time n. On the other hand, the total number of RBs allocated in the first n time slots can at most be the total RBs used to serve all the traffic that arrived within $(0,n]$. Taking the limit as $n\to\infty$ in (10) and the additional assumptions stated in the theorem, it is clear that the time average of $(M_n^\pi)_n$ converges to $\bar{M}^\pi$ and $\bar{M}^\pi = \bar{N}^\pi$. The lower bound in (8) then follows from (7) under the assumptions on arrivals and channels being stationary and independent. □

IV. Simulation results

We consider a BS serving a set of URLLC users and eMBB users. The received SNR was modeled using the 3GPP Urban-Micro path loss model [23], with Rayleigh distributed small scale fading. We use 3GPP Modulation and Coding Scheme (MCS) to find the coding rate C_n per RB based on quantized SNR values in the range -6.934 dB $\leq$ SNR ≤ 20 dB refer [24, Table 5.2.2.1-2]. To determine the channel rate thresholds, we estimate the rate CDF $F_C(\cdot)$ on a given slot for each user using the last 100 channel realizations. Please note that using fewer channel rate samples can lead to errors in CDF estimation, however, we found that the impact on the spectral efficiency was less than 5%. Furthermore, we assume error-free transmissions to users (simulation results for performance under channel transmission errors were not included to space limitation). Finally, 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.

We shall use Guaranteed Rate Scheduler (GRS) as a baseline policy that meets delay deadlines by providing a fixed data rate every time slot. This baseline is analogous to a strict service curve in Deterministic network calculus [8]. One can determine the minimum fixed service rate s for leaky bucket constrained arrivals using the formula [8],

$$s = \frac{\rho\sigma}{(\rho-\mu)d+\sigma}, \tag{11}$$

where ρ is the peak arrival rate, μ is the mean arrival rate, σ is the token buffer size and d is the delay constraint. We also use the multicarrier version of MLWDF [7] policy as a benchmark for scheduling URLLC users.

A. Improvement in eMBB throughput

In this subsection, we demonstrate the throughput improvement for eMBB users where the BS supports 8 different users, 3 URLLC and 5 eMBB users. The five eMBB users are located at distances $520, 560, 650, 720, 800$ meters from the BS. The URLLC users are located at distances $300, 500, 700$ meters from the BS. Users are located at different distances from the BS to capture a realistic, heterogeneous setting. While different settings were considered in terms of the number of users, for simplicity we show results for the 3 URLLC users and 5 eMBB users. We consider ON-OFF bursty arrivals 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. Note that we assume that eMBB users are infinitely backlogged and do not have any stringent QoS requirements, i.e. best effort traffic. Furthermore, we use proportionally fair scheduling [2] at each time slot to select the eMBB user that will be served. The leaky bucket parameters for URLLC users are provided in TABLE I.

A total of 6000 RBs are available to all the users served by the BS, where URLLC users are allocated resources with priority. The RBs that remain unused after allocating resources to all active URLLC users are then used to serve eMBB 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

TABLE I: URLLC user parameters.

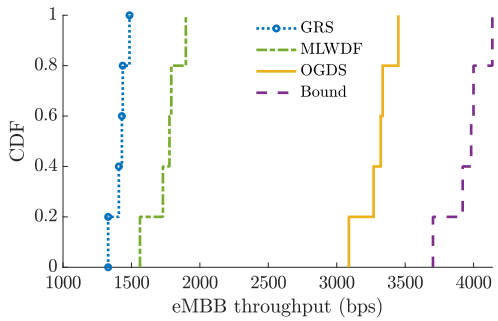


Fig. 2: Long term throughput distribution for eMBB users.

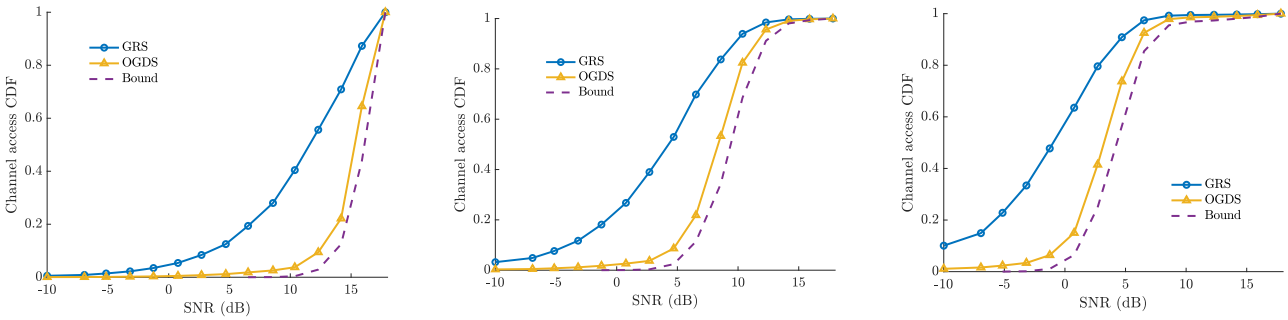
Fig. 2 showcases the throughput gains for eMBB users under various scheduling algorithms for URLLC users. Note that the arrivals (for URLLC users) and the channel rate realizations for all the users are the same while evaluating each of the algorithms. Clearly, OGDS outperforms the baseline GRS policy and the benchmark MLWDF which was designed [7] for QoS provisioning in wireless links. It is indeed very close to the throughput gain bound set by the optimal policy.

B. Improvement in Reliability

Fig. 3 illustrates the improvement in resource allocation by having more transmissions scheduled when the channel rate is higher than the threshold. We plot a weighted distribution of the channel strength per channel use, where the weights are directly proportional to the number of bits scheduled for transmission at that channel strength. This enables us to compare and contrast all the proposed scheduling policies according to the efficiency with which each policy is able to identify the best time slot for data transmission in terms of channel rate. Fig. 3 validates the superior performance of OGDS policy across all users, irrespective of the average channel strength (weak/medium/strong) of the received signal. Furthermore, OGDS policy schedules traffic at a relatively better channel strength – which translates to better reliability as the transmission error rate is a decreasing function of SNR.

C. Network jitter performance

We consider real time video streaming applications to evaluate the jitter performance of our modified OGDS policy. Let traffic arrivals be periodic, with 50 payloads of size 1 KB each that arrive once every 10 milliseconds (ms), over a duration of 10^6 ms for a total rate of 5 Mbps. A range of $4-10$ ms delay deadlines are considered for each payload. Fig. 4(b) exhibits how jitter reduces for each user with higher ζ at the cost of lower spectral efficiency as shown in Fig.



(a) Strong user. (b) Medium user. (c) Weak user.

Fig. 3: The weighted distribution of channel strength per channel use for data transmission under all three policies.

4(a). The plots for modified OGDS are labeled as "ζ-OGDS", where ζ is the jitter control parameter. Also, note that as the delay deadline increases we see a fall in the total number of RBs required for all policies.

D. Spectral efficiency under nonstationary environment

Our design of the dynamic threshold that triggers packet scheduling was based on the assumption that channel rate variations are i.i.d across time slots. However, in practice, wireless channel rates are often nonstationary, depending on user mobility and other propagation dynamics. To evaluate the performance of our scheduler on non-stationary wireless channels we used a trace [18] driven simulation for all three policies. In our simulation, we used 15 samples of past channel realizations to track the empirical CDF of the wireless channel. Fig. 4(c) demonstrates that OGDS is very close to the offline lower bound in a practical real world wireless environment.

E. Admission Control

We consider a set of 100 users with ON-OFF bursty traffic as described in IV-B with parameters (in packets per time slot) $\sigma = 50, \rho = 10, \mu = 5$. Each user's delay deadline and distance from the BS were drawn uniformly random from their respective range of values as shown in TABLE II. The channel and traffic dynamics were generated over 10^6 time slots. We denote the total resource requirement of users $1, \ldots, u$ that are admitted into the system by the random variable X_u.

Parameter	Range of values
Distance	$\{200, 250, \ldots, 800\}$
Delay Deadline	$\{2, 3, \ldots, 10\}$

TABLE II: URLLC user delay and traffic parameters sample space for simulation.

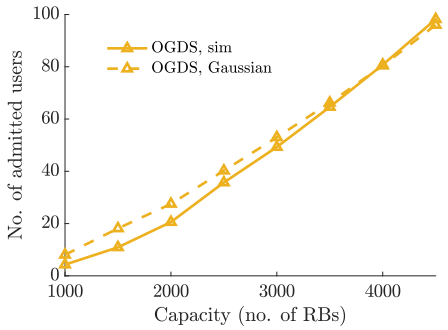
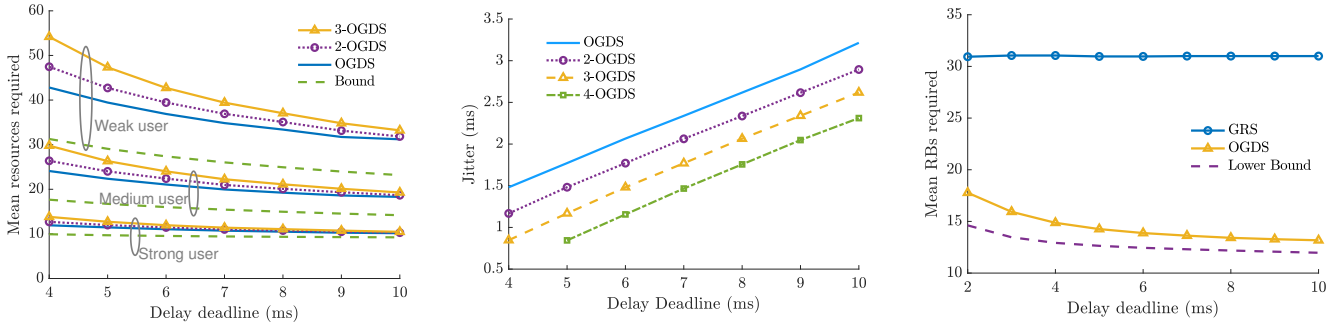


Fig. 5: Admission control using large deviation bounds on the total RBs required for all admitted URLLC users.

Fig. 5 shows the number of users that can be admitted into



(a) Spectral efficiency. (b) Jitter performance. (c) Non-stationary environment.

Fig. 4: Spectral efficiency, packet jitter of a single user as a function of the delay and jitter control parameter ζ.

the system as a function of the system capacity $\overline{m}$. The solid lines denote the number of users u that could be admitted such that the probability of X_u exceeding $\overline{m}$ is less than $\delta = 10^{-3}$. For the same set of users, we also use the Gaussian approximation $X_u \sim \mathcal{N}(\mu_u, \sigma_u^2)$, where the aggregate mean and variance, denoted $\mu_u = \sum_{i=1}^{u} \mu_i, \sigma^2 = \sum_{i=1}^{u} \sigma_i^2$, are directly measured by observing the resource allocation to admitted users under OGDS. Let $Y \sim \mathcal{N}(\hat{\mu}, \hat{\sigma}^2)$ model the random resource requirement for a new user. The probability that the total resource requirement $X_u + Y$ will exceed $\overline{m}$ is approximated using the following inequality, see [25],

$$\mathbb{P}\left(X_u + Y > \overline{m}\right) \leq e^{-\frac{(\overline{m}-\mu)^2}{2\sigma^2}}, \tag{12}$$

where $\mu = \mu_u + \hat{\mu}$ and $\sigma^2 = \sigma_u^2 + \hat{\sigma}^2$. Note that the inequality in (12) provides a computationally reasonable expression that can be used to decide if the new user can be admitted without exceeding δ.

Typically, Y is unknown, so we use the mean $\hat{\mu} = \frac{1}{u}\mu_u$ and variance $\hat{\sigma}^2 = \frac{1}{u}\sigma_u^2$ as a proxy for a typical new user. Note that one could also use the worst case user statistics (of currently admitted users) as a proxy for the new user, specifically, $\hat{\mu} = \max_{1 \leq i \leq u} \mu_i$ and $\hat{\sigma}^2 = \max_{1 \leq i \leq u} \sigma_i^2$. It can be seen in Fig. 5 that as long as the system capacity is large enough, one can use Gaussian approximation to model the aggregate resource requirement for admission control.

V. Conclusion

We have developed a new measurement based opportunistic scheduler for delay and jitter constrained traffic and demonstrated its efficiency by showing how close its spectral efficiency is to the optimal offline scheduler. An optimal offline scheduler can be designed if complete information on future channels is known with high accuracy. One could consider applying machine learning techniques to predict future channels to help reduce the efficiency gap (with respect to the optimal). While state-of-the-art ML predictors are accurate over a smaller horizon, research on accurate prediction over a much longer time scale is still in progress, especially when the wireless environment is non-stationary. Finally, we have also demonstrated the *robustness* of our proposed scheduler in tracking past channel variations to identify good transmission opportunities despite nonstationary wireless channel variations. An interesting future research direction is to model user mobility and identify the optimal window of past channel samples to best estimate the empirical CDF of the user.

References

[1] A. L. Stolyar, "Maximizing Queueing Network Utility Subject to Stability: Greedy Primal-Dual Algorithm," *Queueing Systems*, vol. 50, pp. 401–457, 2005.

[3] 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.

[2] 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] 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.

[5] S. Shakkottai and A. L. Stolyar, "Scheduling for Multiple Flows Sharing a Time-varying Channel: The Exponential Rule,"

[6] 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.

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

[8] 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.

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

[10] 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.

[11] 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.

[12] 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.

[13] 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.

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

[15] 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.

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

[17] 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.

[18] R. Hernangomez, P. Geuer, A. Palaios, D. Schäufele, C. Watermann, K. Taleb-Bouhemadi, M. Parvini, A. Krause, S. Partani, C. Vielhaus, M. Kasparick, D. F. Külzer, F. Burmeister, S. Stanczak, G. Fettweis, H. D. Schotten, and F. H. P. Fitzek, *Berlin V2X*. IEEE Dataport, 2022.

[19] G. Grimmett and D. Stirzaker, *Probability and Random Processes*, vol. 80. Oxford university press, 2001.

[20] S. Patil *et al.*, "Measurement-based Opportunistic Scheduling for Heterogenous Wireless Systems," *IEEE Trans. Commun.*, vol. 57, no. 9, pp. 2745–2753, 2009.

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

[22] 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.

[23] 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.

[24] 3GPP, "5G NR physical layer procedures for data," Technical Specification (TS) 38.214, 3rd Generation Partnership Project (3GPP), 04 2018. Version 15.2.0.

[25] A. Dembo and O. Zeitouni, *Large Deviations Techniques and Applications*. Stochastic Modelling and Applied Probability, Springer Berlin Heidelberg, 2009.