Generalized Performance Analysis and Optimization of Enhanced Power Saving Semi Persistent Scheduler

Prashant K. Wali

Abstract: **Enhanced-power saving-semi persistent scheduler (E-PS-SPS) for downlink voice over LTE (VoLTE) traffic was proposed to reduce the energy consumption of the LTE eNodeB. Further, it was established that E-PS-SPS outperforms traditional SPS (T-SPS) both in terms of energy consumption and capacity (maximum number of simultaneous calls it can handle). However, in order to establish the superiority of E-PS-SPS, the analysis and comparison was done with two strong simplifying assumptions; (*i*) same downlink mean signal-to-noise ratio (SNR) for all the VoLTE users even though the mean SNRs are independently and identically distributed (i.i.d.) and (*ii*) same instantaneous downlink channel gains on different physical resource blocks (PRBs) resulting in same instantaneous SNR on all allocated PRBs for a user in a packet transmission attempt, even though channel gains on different PRBs are i.i.d.. In this paper, we carry out a more general analysis of E-PS-SPS in a more realistic scenario by removing these two strong assumptions and derive a closed-form expression for its capacity and achievable success rate. The analytical expressions are validated using Monte Carlo simulations. We then use these expressions to show how the performance of E-PS-SPS can be maximized for a desired success rate. Finally, we also show through extensive simulations that the capacity of E-PS-SPS is same as T-SPS in the realistic scenario which we consider in the work. However, T-SPS has a marginally better success rate than E-PS-SPS which can be overlooked considering the significant energy saving that E-PS-SPS can offer.**

Index Terms: **Capacity, enhanced power saving semi persistent scheduler, optimal performance, success rate, VoLTE, .**

I. INTRODUCTION

THERE has been a substantial attention paid to design mechanisms/algorithms in order to arrest the energy consumption of mobile broadband wireless technologies like 4G LTE, owing to their large scale deployment. Studies have revealed that base stations consume upto 60% of the total power in these technologies while the remaining power gets divided between other components. Since the power amplifier (PA) is identified as the largest consumer of power within a base station, with a share of around 65% of the total input power [1], a lot of research has been directed at bringing down the energy consumption of the base stations by focusing on the PA.

Manuscript received October 12, 2020; revised January 31, 2021; approved for publication by Tony Quek, Division II Editor, March 3, 2021.

This research was supported by BITS-Pilani, Hyderabad campus.

The author is with the Department of Electrical and Electronics Engineering, BITS-Pilani, Hyderabad Campus, India., email: iwalihere@gmail.com.

Prashant K. Wali is the corresponding author.

Digital Object Identifier 10.23919/JCN.2021.000008

In addition to improving their efficiency through better designs, energy saving mechanisms that target the PA can also be based on enabling their sleep modes, since putting a PA into sleep mode reduces its power consumption drastically [2]. Hence, PA can be opportunistically pushed into sleep modes at the link level (link between the base station and user equipment (UE)) through intelligent traffic scheduling techniques in the downlink to reduce the base station energy consumption [1], [3], [4]. Even though a deep sleep mode can bring down the power consumption of a PA to a negligible value, it has been reported that using the light sleep mode is beneficial in order to have a small wake up time of the PA (because of the stringent requirements of broadband technologies) [2]. Even though the energy saving in a light sleep mode is lesser than deep sleep mode, the wake up of the PA time is drastically reduced avoiding the adverse effect on the operation of the cell [2].

In addition to the aforementioned references, some of the other works that have investigated the benefits of introducing sleep modes of the PAs include [5], [6]. The idea of intelligent traffic scheduling to enable sleep modes of the PA was further exploited in [7], in which the authors proposed a power saving-semi persistent scheduler (PS-SPS) to reduce the energy consumption of LTE base station (called as eNodeB). PS-SPS introduces a novel modification to the traditional semi-persistent scheduler (T-SPS) resource allocation rule for the VoLTE packets in the downlink of a LTE cell. The authors subsequently proposed enhanced-PS-SPS (E-PS-SPS) which further improved the resource allocation rule for downlink VoLTE packets to increase the energy savings of the LTE eNodeB [8]. The authors also showed that the capacity[1] of E-PS-SPS is better than PS-SPS and upper bounds the capacity of T-SPS. This is a consequence of the novel resource allocation rule that E-PS-SPS employs for downlink VoLTE packets as compared to PS-SPS and T-SPS. E-PS-SPS allocates resources for voice over LTE (VoLTE) packets in a structured way such that a balance is created between the number of physical resource blocks (PRBs) required for fresh transmissions and re-transmissions, while simultaneously ensuring that the number of physical downlink control channel (PDCCH) control signals required does not cross the available limit. E-PS-SPS allocates same number of downlink PRBs for every user's VoLTE packets. This is done in order to simplify its implementation and computational complexity. However, the performance analysis in [8] was done with two very strong assumptions. Firstly, same downlink mean signal-to-noise ratio (SNR) was assumed for every user despite users' differing distance and shadowing from the eNodeB which

[1] We will define the term capacity more accurately in a later section.

should actually result in their downlink mean SNRs being i.i.d. The mean SNR that was assumed for all the users ensured that the effective instantaneous received SNR of every packet reaches a threshold SNR that is required for the packet to be decoded correctly by the UE[2] with probability approaching 1, after maximum number of possible transmissions within the delay constraint. Secondly, even though the instantaneous channel gains across different PRBs are i.i.d., they were assumed to be same on all the allocated PRBs for a user. Even though this kind of analysis did not introduce any loss of generality in proving the superiority of E-PS-SPS over PS-SPS and T-SPS, it does not provide the performance of E-PS-SPS in a realistic scenario in which these two strong assumptions do not hold true. Additionally, the key problem of optimizing the performance of E-PS-SPS in such a scenario still remains to be addressed. In this regard, the main contributions of this work can be summarized as follows:

1. We carry out performance analysis of E-PS-SPS and derive a closed form expression for its capacity in a general setting in which VoLTE users have i.i.d. downlink mean SNRs and i.i.d. instantaneous SNRs on their allocated PRBs.
2. We provide an expression to calculate the success rate (percentage of satisfied users among the total connected) that E-PS-SPS can enable.
3. After validating our analytical expressions with Monte Carlo simulations, we show that given a desired success rate, how these expressions can be used to maximize the performance of E-PS-SPS, i.e., the optimal number of PRBs to allocate to users, in order to maximize the capacity.
4. Finally, we show through extensive simulations that the capacity of E-PS-SPS is same as T-SPS in the realistic scenario which we consider in this work. However, T-SPS has a marginally better success rate than E-PS-SPS which can be overlooked considering the significant energy saving that E-PS-SPS can offer.

Since LTE has a large scale deployment and will remain as a complementary technology for a long time, as outlined in the non-standalone (NSA) and standalone (SA) version of 3GPP NR in the proposed road-map towards 5G [9], [10], SPS in LTE has been getting considerable attention even to this day signifying its usefulness. Additionally, SPS or a variant of it designed for LTE will remain relevant since it can also be compatible with new radio (NR) technology developed by 3GPP, as both technologies share a similar medium access control (MAC) layer skeleton [10]. Semi persistent scheduling has also been considered for Internet of things (IoT) applications as shown in [10]–[12]. In [13], a sensing-based SPS is analyzed for its performance in C-V2X networks. An efficient semi-persistent scheduling algorithm capable of allocating heterogeneous periods and supporting multiple platoons is presented in [14]. The authors in [15] present a method for latency reduction in narrowband LTE with SPS. In [16], a priority based SPS is proposed for VoLTE. More recently, the authors in [17] indicate the importance of continued analysis of SPS by investigating the control overhead reduction brought about by SPS. In this regard, our contributions summarized above can also be considered as useful since they aid us in obtaining a significantly more crucial and detailed insight into the working of E-PS-SPS as compared to the work in [8]. To the best of our knowledge, this kind of a work is missing in the literature.

[2]The terms user and UE will be used interchangeably henceforth.

The rest of the paper is organized as follows. In Section II, we first briefly recall the LTE physical layer frame format. Then, in order to help the reader with a quick background of the work about E-PS-SPS in [8], we provide a brief description of T-SPS, PS-SPS, and E-PS-SPS. The analysis of E-PS-SPS is carried out in Section III, in which we derive capacity and success rate expressions for the realistic scenario considered in this paper. In Section IV, we validate our analytical expressions using Monte Carlo simulations. We then describe how to maximize the performance of E-PS-SPS for a desired success rate using the results obtained through our analytical expressions. Further, we also compare the performance of E-PS-SPS with T-SPS through extensive simulations. Section V concludes the paper.

II. BACKGROUND AND ENHANCED-POWER SAVING-SPS

In this section, we provide a quick and brief background of LTE physical layer frame structure, T-SPS, PS-SPS and E-PS-SPS. For a more elaborate description of these topics, please see [8] and references therein.

A. LTE Physical Layer Frame Structure

Each downlink frame in LTE is of 10 ms duration and made up of ten 1 ms subframes, each of which are divided into two 0.5 ms slots, with each slot containing seven OFDM symbols. The total operating bandwidth is divided into subcarriers, spaced 15 kHz apart. A subcarrier with a 15 kHz bandwidth along frequency and 1 OFDM symbol duration along time is called a resource element (RE). A set of 12 consecutive subcarriers for a duration of 1 ms (14 symbols) is called a PRB and contains 168 REs. A PRB is the smallest unit of resource that can be allocated to a user during traffic scheduling. The number of PRBs in one subframe depends on the total bandwidth. For instance, 5 MHz cell would have 25 PRBs. The eNodeB includes downlink and uplink scheduling grants in PDCCH (limited in number), which are used by the UEs to know in which uplink/downlink PRBs they have to send and/or receive data in a scheduling interval. In LTE, traffic scheduling is performed every milli second which is the duration of one subframe [18], [19].

B. T-SPS

In LTE downlink traffic scheduling, each PRB in a subframe is allocated to one of the UEs based on their channel feedback and the QoS requirement. This allocation is intimated to UE through a PDCCH signal. Each UE decodes the PDCCH signal to know if the scheduler has scheduled it in the current subframe and if so, in which PRB. This is called dynamic scheduling. Dynamic scheduling if used for VoLTE users (packets) will require plenty of PDCCH signaling because of the small packet sizes and constant packet inter-arrival time of 20 ms [20].

With a limited number of PDCCH signaling resources in a subframe, it becomes difficult to support large number of VoLTE users in the system if dynamic scheduling is employed. Hence semi-persistent scheduler (SPS) was proposed. In SPS, each UE is persistently allocated PRBs where initial transmissions of freshly arrived packets can be done without PDCCH signaling. If a first transmission of this packet fails, the scheduler tries to dynamically schedule the re-transmission of the failed packet (using hybrid ARQ (HARQ)) by using PDCCH signal [20]. We will call this form of SPS as traditional-SPS (T-SPS).

Once a VoLTE call from a UE is admitted and the first packet arrives for this UE, starting from and including the subframe in which the packet arrived, T-SPS waits for a subframe which has a free PDCCH to signal and sufficient PRBs to allocate to this UE. Since voice packets arrive (or are sent) every 20 ms for/from the UE, T-SPS allocates (dedicates) the same set of PRBs for this UE in every 20th subframe (since each subframe is of 1 ms interval) from then on, i.e., the UE is allocated a fixed set of PRBs to send and/or receive its VoLTE packets in every 20th uplink and/or downlink subframe and informed of this allocation through PDCCH signal sent for the first packet. Therefore, the initial transmission of subsequent packets that arrive every 20 ms only use those dedicated PRBs in every 20th subframe but do not need PDCCH signaling. Observe that since PRBs are dedicated in every 20th subframe for each active UE, we can treat the downlink subframes as a sequence of cycles in time, with each cycle containing 20 subframes numbered 1–20, since the PRBs allocation repeats every 20 subframes. Hence, one can picturize the PRBs allocation to active UEs (or active calls) by observing any set of 20 subframes along time.

Fig. 1(a) shows an example of PRBs dedicated by T-SPS as six VoLTE calls arrive and 3 PRBs are used for each VoLTE packet. Observe that since these calls arrive randomly in time and T-SPS looks for the earliest subframe having free PDCCH signal and sufficient free PRBs to allocate to each call from the time of its arrival, the PRBs allocation also is randomly spread across the cycle. Hence, we can see from Fig. 1(a) that only a fraction of the total available PRBs might get dedicated and scheduled in each subframe in which initial downlink transmission of VoLTE packets happens.

C. PS-SPS

PS-SPS was proposed as a modification of T-SPS for reducing the energy consumption of the eNodeB because of the following observation.The power consumption of a LTE eNodeB is seen to grow linearly with increase in the number of PRBs on which data is scheduled in the downlink [2], [3]. However, it is offset by a fixed constant value. The power P_C consumed to transmit data on n PRBs can be modeled as [3]:

$$P_C = P_f + n\beta P_{PRB}, \tag{1}$$

where, β is the loss factor of the PA. The fixed offset power P_f is a large fraction of the total power consumed by the PA in ON state and is independent of the number of transmitted PRBs in a subframe. Hence, PA can be switched to light sleep mode when there is no data to be transmitted in a subframe [2], [3] reducing its power consumption to a small fraction of P_f (and hence P_C). Hence, PS-SPS is motivated by the following two observations [8].

(i) The traffic load has negligible influence on instantaneous power consumption of a macro eNodeB in subframes in which data is scheduled on PRBs (called as *active subframes* in this work).
(ii) The components of a transmitter section can be put into idle sleep mode in those subframes in which there is no data to be transmitted. The power consumption of the transmitter section in idle sleep mode is significantly less compared to its active state.

Owing to the aforementioned observations, PS-SPS was proposed to transmit maximum downlink VoLTE traffic (as many VoLTE packets as possible) in active subframes, so that the number of active subframes required to schedule VoLTE traffic is minimized and the number of sleeping subframes is maximized.

To achieve the above objective, PS-SPS treats the downlink subframes as a sequence of cycles in time, with each cycle containing 20 subframes numbered 1–20 as mentioned in previous subsection. As calls arrive, unlike T-SPS which looks to allocate the first set of free PRBs starting from the subframes in which the calls come in the cycle, PS-SPS looks to allocate the first set of free PRBs starting from subframe 1 in the cycle for initial transmissions of packets, irrespective of when each admitted call arrives. Hence, it allocates all the PRBs from subframe 1 first, then from subframe 2 till all its PRBs are fully utilized and follows the same rule for every subsequent subframe. Hence, it allocates PRBs to a call from subframe n in the cycle only if all the PRBs of subframes 1 to $n-1$ are already allocated. In other words, whenever a new call arrives and is admitted, irrespective of in which subframe between 1–20 that call arrived within the cycle, PS-SPS sequentially looks for a subframe with sufficient free PRBs beginning from the 1st subframe in the cycle, to allocate to this call. This allocation is signaled to the UE in the subframe in which it finds a free PDCCH. As an example, let us assume that a call arrives in the 10th subframe of a cycle. Let us also assume there are some free PRBs in subframe 2 sufficient to be allocated to this call and all PRBs are free in subframes 15 to 20 in the cycle with all PRBs in other subframes already allocated. Since T-SPS waits for a subframe with sufficient resources from subframe 10, it allocates PRBs in subframe 15 making it an active subframe. But since PS-SPS searches from subframe 1 in the cycle for free PRBs, it would therefore allocate the free PRBs from subframe 2, thus retaining the sleep mode of the eNodeB in subframe 15 unlike T-SPS.

Because of the resource allocation rule as indicated above, PS-SPS is expected to increase the resource utilization of each active subframe by allocating PRBs till they are all utilized. This minimizes the number of active subframes (and maximizes the number of sleeping subframes) compared to T-SPS. Fig. 1(b) shows how PS-SPS would dedicate PRBs to the same 6 active UEs as compared to T-SPS shown in Fig. 1(a). It can be seen from Figs. 1(a) and 1(b) that PS-SPS dedicates all PRBs in the first two subframes in each cycle for the 6 active VoLTE users, while for T-SPS, the dedicated PRBs might be spread randomly. In case of re-transmissions, PS-SPS handles them like T-SPS through dynamic scheduling.

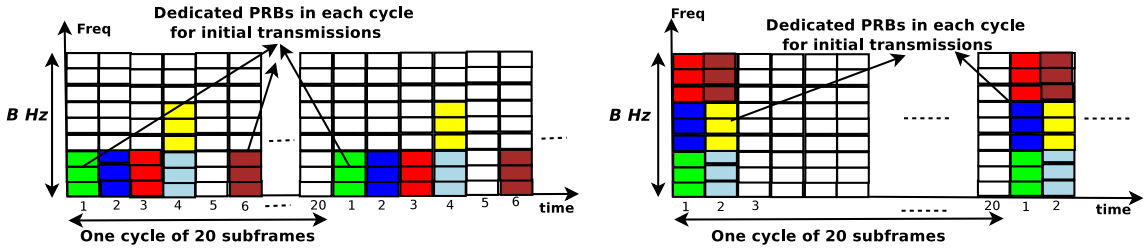


Fig. 1. PRB Allocation by T-SPS and PS-SPS for 6 Active UEs. 3 PRBs used for every VoLTE packet. PRBs allocated for each call is represented by different colour. Observe that PS-SPS uses only two subframes in each cycle to support 6 active calls unlike T-SPS which might use more: (a) T-SPS allocation and (b) PS-SPS allocation.

D. E-PS-SPS

Following the same objective as PS-SPS, E-PS-SPS also employs a novel resource allocation policy for scheduling downlink VoLTE packets to utilize as less a number of subframes as possible so that maximum number of subframes can be left free for enabling PA sleep mode [8]. We provide a brief description of how it achieves this objective below.

Since VoLTE packets arrive every 20 ms for a user in listening state (state in which the user is listening to the other party speaking), a fixed set of PRBs in every 20th downlink subframe (since each subframe is of 1 ms interval) needs to be allocated. Hence, E-PS-SPS treats the downlink subframes as cycles containing subframes 1–20, in time. Observe that because calls arrive randomly in time, with T-SPS, the PRBs allocation would be randomly spread within the 20 subframes cycle. This results in only a fraction of the total available PRBs getting dedicated and scheduled in each subframe in which initial (first) downlink transmission of freshly arriving VoLTE packets occurs.

However, unlike T-SPS, which looks to allocate free PRBs starting from the subframe in which a call arrives, E-PS-SPS allocates free PRBs to users starting from subframe 1 in the cycles for initial (first) transmission of packets, irrespective of the subframe in which a call arrives. For example, if the first call arrives in subframe 5 of the current cycle, its packet is allocated PRB resources in subframe 1 of next cycle provided its maximum quota is not yet reached[3]. Every subsequent packet that arrives in subframe 5 (packets arrive every 20 ms) for this call gets scheduled in subframe 1 of next cycle. However, not all PRBs in a subframe are allocated for initial transmissions. Some PRBs are left free to accommodate dynamically scheduled re-transmissions also. Once PRBs have been allocated for the maximum quota of initial transmissions in subframe 1, E-PS-SPS then starts allocating PRBs from subframe 8 of the cycle as more calls arrive. This is done for the following reasons. In LTE-A, it takes a total of 7 ms for the scheduler to know about the failure of a packet transmission and preparing its re-transmission [19]. Hence the scheduler can attempt a re-transmission starting only from the 8th subframe after the previous transmission. Therefore, notice (see Fig. 2) that if the initial transmission of a packet in subframe 1 from the current cycle fails, its 1st and 2nd re-transmissions can be attempted in subframe 8 and 15, respectively, in the same cycle and 3rd re-transmission in subframe 2 of next cycle and so on[4], after each failed attempt, till the delay threshold of 50 ms is violated. Observe that the 1st re-transmission of a packet is more likely than the subsequent ones. Therefore, subframe 8 has more chance to be used for 1st re-transmissions of failed packets than subframe 2 for 3rd in each cycle when subframe 1 has initial transmissions. Hence, if new calls arrive after subframe 1's maximum quota is reached, it is more efficient to allocate PRBs from subframe 8 next. Note that this form of organized resource allocation minimizes the number of subframes used to handle the VoLTE calls (packets). This is in contrast to T-SPS which does not follow any pattern for resource allocation and ends up unnecessarily making use of more than required under-utilized subframes.

Continuing in the same manner, subframe 15 is chosen after subframe 8's quota is used up, followed by subframe 2 and so on, to allocate PRBs for initial transmissions of VoLTE packets as new calls arrive. Hence, the sequence that E-PS-SPS follows in allocating the subframes for initial transmissions of VoLTE packets is given by S_n = 1, 8, 15, 2, 9, 16, 3, 10, 17, 4, 11, 18, 5, 12, 19, 6, 13, 20, 7, and 14. Hence, each subframe when used will have the following kind of packets to be transmitted:

1. Initial transmissions (fresh packets).
2. 1st re-transmission of those packets that had initial transmission 8 subframes earlier but were not successful.
3. 2nd re-transmission of packets that had initial transmission 16 subframes earlier but were not successful in both their initial transmission and the 1st re-transmission 8 subframes later.
4. 3rd re-transmission of packets that had initial transmission 24 subframes earlier but were not successful in their initial transmission, and also in 2 re-transmissions that were successively spaced 8 subframes apart.
5. 4th re-transmissions of packets that had initial transmission 32 subframes earlier but were not successful in their initial transmission, and also in all the 3 re-transmissions that were successively spaced 8 subframes apart.

[3]Maximum quota will be derived in Section III to obtain the capacity of E-PS-SPS.

[4]If sufficient PRBs and a PDCCH resource is available in each of these subframes for re-transmission.

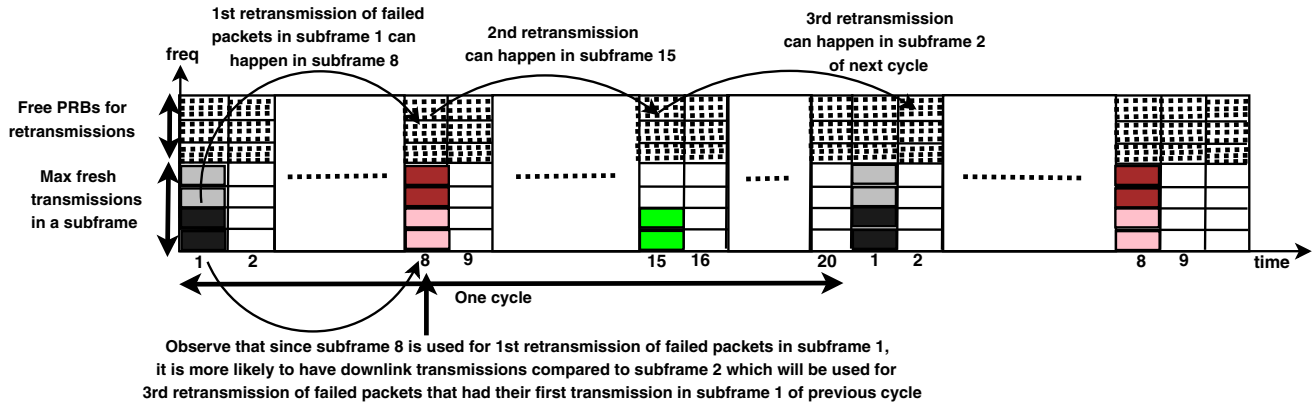


Fig. 2. E-PS-SPS resource allocation algorithm.

Table 1. Summary of the symbols used.

Symbol	Parameters
$\mathcal{M}$	Number of PRBs dedicated (once in every 20 subframes) to the initial transmission of packets corresponding to an admitted VoLTE user in the listen state.
$\mathcal{L}$	Size of each VoLTE packet in bits
$\mathcal{N}_{prb}$	Number of PRBs available in each subframe
$\mathcal{S}_i$	Set of link adaptation SNR thresholds in LTE
r_i	Rate achievable in bits/symbol, if the received SNR $\in [T_{i-1}, T_i)$
$\gamma_{k,n}$	Instantaneous received SNR of the nth transmission at user k
μ_i	Downlink mean SNR for UE i
$\gamma_{k,n}^{eff}$	Instantaneous effective SNR at UE k after n transmissions using Chase combining technique
$\mathcal{N}_{\max}$	Maximum number of possible transmissions within the delay threshold of a VoLTE packet
$\mathcal{N}_{vp}$	Maximum number of VoLTE packets a subframe can hold
$\mathcal{N}_{pd}$	Number of PDCCH resources in each subframe to indicate downlink grants

III. PERFORMANCE ANALYSIS OF E-PS-SPS

A. System Model

We highlight the main aspects of the system model before proceeding to the analysis. Some of the frequently used symbols and their meanings are indicated in Table 1 for ease of reference.

A LTE cell with $\mathcal{N}_{prb}$ downlink PRBs is considered. Each VoLTE user in listen state is allocated $\mathcal{M}$ PRBs. This means that $\mathcal{M}$ PRBs are persistently allocated in every 20th subframe for the user as long as the user is in listen state. A VoLTE packet transmitted in the downlink is considered successful if the effective instantaneous received SNR exceeds the required threshold SNR at the user's UE. If not, the scheduler can keep re-transmitting the packet till it is successful or its delay threshold is violated, making it a failure and getting dropped. As already mentioned, a re-transmission of a previously failed attempt can be tried only from the 8th subframe from the time of its previous transmission, provided the subframe has enough free PRBs and a PDCCH for signaling the dynamic re-transmission to the UE. Else, the scheduler has to wait for such a subframe. Let the maximum number of VoLTE packets (fresh and re-transmissions put together) that each subframe can hold be $\mathcal{N}_{vp}$. Let $\mathcal{N}_{pd}$ be the number of PDCCH signaling resources available in each subframe to indicate downlink scheduling grants when required. We assume chase combining technique for HARQ re-transmissions [21]. Therefore, the effective instantaneous received SNR for a packet seen by a UE after n transmissions is the total of the instantaneous received SNRs in each of its transmission attempts including the nth transmission.

We will call the $\mathcal{M}$ allocated PRBs to a user as a *scheduling block*. A Rayleigh fading channel that is constant over a 1 ms subframe (coherence time of 1 ms in LTE) with all the subcarriers within a PRB having the same channel gain (coherence bandwidth of 180 KHz in LTE) is considered [22]. Hence, the received signal in the ith PRB within a user k's scheduling block is given by,

$$y_{i,k} = h_{i,k}x_{i,k} + w_{i,k}, \ \ \forall 1 \leq i \leq \mathcal{M},$$

where, $x_{i,k}$ is the transmitted signal (or symbol) on the ith PRB in the scheduling block, $h_{i,k}$ is the downlink channel gain over the ith PRB and $w_{i,k}$ is circular symmetric complex zero mean Gaussian noise with unit variance. For Rayleigh channels, $h_{i,k}$ is also a circular symmetric Gaussian random variable with its variance depending on the distance of the UE from the eNodeB and shadowing. Hence, the instantaneous received SNR by the UE k on ith PRB will be equal to $\gamma_{i,k} = |h_{i,k}|^2$. Therefore $\gamma_{i,k}$ is an exponential random variable with mean $\mu_{i,k}$ which is equal to the sum of the variances of the real and imaginary parts of $h_{i,k}$ [23]. This is unlike [8] which assumes instantaneous received SNRs to be the same across all PRBs of a user's scheduling block. Since the mean SNR for a user is constant across the entire bandwidth [24], we will denote $\mu_{i,k}$ as μ_k signifying its independence on PRB index i. We consider users to be uniformly spread across the cell and hence model their mean SNRs as distributed uniformly over a closed interval $[a, b]$. This is also unlike [8] in which all the users are assumed to experience the same mean SNR. Hence, different from [8] in these two important aspects, this work models instantaneous SNRs across the PRBs within a user k's scheduling block as i.i.d. exponential random variables with mean μ_k that is uniformly distributed over a closed interval $[a, b]$. However, the mean SNR is known to remain constant in time for at least few seconds [25].

The set of link adaptation SNR thresholds are denoted as $\mathcal{S}_0, \mathcal{S}_1, \cdots, \mathcal{S}_L$ and determine the maximum rate in bits/RE that can be achieved in downlink transmission. These thresholds help achieve a target block error rate of less than 10% [19, Fig. 10.1], should the eNodeB transmit over the allocated

scheduling block to a user i at a rate r_j bits/RE, if the instantaneous received SNR $\gamma_i \in [\mathcal{S}_{j-1}, \mathcal{S}_j)$ on all PRBs. A block error rate of less than 10% ensures that the packet is decoded correctly as done in [26]. Each rate r_j corresponds to one of the 16 modulation and coding schemes (MCS) that the eNodeB can employ for transmission of data in the downlink. This list is provided in [19, Table 10.1]. We denote the effective instantaneous SNR of the received packet after n transmissions (i.e., $n-1$th re-transmission after the first transmission) for PRB i in user k's scheduling block as $\gamma_{i,k,n}^{eff}$. Hence, a VoLTE packet transmitted at rate r_j (which depends on the number of PRBs $\mathcal{M}$ allocated to users) by the scheduler to a UE k is decoded correctly by the UE in the nth transmission if $\gamma_{i,k,n}^{eff}$ is at least equal to $\mathcal{S}_{j-1}$ on all PRBs of the scheduling block, i.e., $\gamma_{i,k,n}^{eff} \geq \mathcal{S}_{j-1}, 1 \leq i \leq \mathcal{M}$.

The number of data carrying REs in a PRB is denoted as D. The size of each VoLTE packet be V bits. With a fixed size scheduling block of $\mathcal{M}$ PRBs for each user, the resulting rate of packet transmission for each user is $r = V/\mathcal{M}D$ bits per Resource Element. For this rate r, we denote the threshold SNR required on each PRB of a scheduling block for the packet to be decoded correctly by the UE as $\mathcal{S}_{th}$. The threshold delay for the VoLTE packet to reach the UE is set to 50 ms [8], [27]. We fix the maximum number of transmissions possible (if required) for each packet at 5 within this 50 ms constraint [8].

B. E-PS-SPS Capacity Analysis

The capacity of E-PS-SPS in a general setting (as described in the system model) is defined as the maximum number of users for whom the scheduler can enable maximum number of possible transmissions (if required) for each packet within the delay constraint [8]. Hence,

$$\mathcal{C} = \max\{\mathrm{n} : \mathrm{P}\left(\mathcal{N}_\mathrm{i} = \mathcal{N}_\mathrm{max}\right) = 1, \quad \forall \mathrm{i} \leq \mathrm{n}\}, \tag{2}$$

where, $\mathcal{N}_\mathrm{max}$ denotes the maximum number of transmissions achievable within a packet's delay threshold and $\mathcal{N}_i$ denotes the number of transmissions of a packet of user i that the scheduler can achieve within the delay constraint.

As done in [8], the capacity of E-PS-SPS in a general setting will also be obtained by finding the maximum number of fresh packets $\mathcal{F}_i^{\max}$ that can be filled in each of the subframes in a cycle, such that all these packets can be scheduled $\mathcal{N}_\mathrm{max}$ times (if needed) within their delay constraint. Since E-PS-SPS fills each subframe with the same number of fresh packets, we denote $\mathcal{F}_i^{\max}, i = 1, 2, \cdots, 20$ as $\mathcal{F}^{\max}$ for all subframes henceforth.

In order to derive capacity, we have to now find $\mathcal{F}^{\max}$ under the following two constraints for each subframe in a cycle:

- ***Constraint 1:*** Since a packet has 40 ms on an average before its violation of the delay constraint [8], it can keep getting re-transmitted up to 4 times every 8th subframe if its initial and subsequent re-transmissions keep failing (since the 5th re-transmission cannot happen before 40 ms and violates the delay constraint). Hence, in each subframe we can consider only those packets for re-transmission that had their initial transmission within the previous 32 ms. For example, in subframe 1 of a cycle, we can consider only the following packets for re-transmission:
  1. Packets that had a failed initial transmission 8 subframes earlier (i.e., subframe 14 of previous cycle) and are ready for 1st re-transmission now. Let the number of such packets be denoted as R_1^{14}.
  2. Packets that had initial transmission 16 subframes earlier (i.e., subframe 7 of previous cycle) but failed in both their initial transmission and the 1st re-transmission 8 subframes later and hence are ready for 2nd re-transmission now. Let the number of such packets be denoted as R_2^7.
  3. Packets that had initial transmission 24 subframes earlier (i.e., subframe 20 in a couple of cycles before the current one) but failed in their initial transmission, and also in 2 re-transmissions that were successively spaced 8 subframes apart and are ready for 3rd re-transmission now. Let the number of such packets be denoted as R_3^{20}.
  4. Packets that had initial transmission 32 subframes earlier (i.e., subframe 13 in a couple of cycles before the current one) but failed in their initial transmission, and also in all the 3 re-transmissions that were successively spaced 8 subframes apart and are ready for 4th re-transmission now. Let the number of such packets be denoted as R_4^{13}.

  However, it is also to be noted that in each subframe, there are a maximum of $\mathcal{N}_{pd}$ PDCCH signaling resources for re-transmissions. Therefore, the expected number of total re-transmissions in each subframe should not exceed $\mathcal{N}_{pd}$. Hence, considering the above observations, we can express the first constraint for subframe 1 as:

$$\mathbb{E}\left[R_4^{13} + R_3^{20} + R_2^7 + R_1^{14}\right] \leq \mathcal{N}_{pd}$$

  Following a similar line of argument for each subframe, we can express *Constraint 1* for subframes $1-20$ (denoted as SF-1 to SF-20) through the following set of equations as shown below in (3).

$$\begin{aligned}
\mathbb{E}\left[R_4^{13} + R_3^{20} + R_2^7 + R_1^{14}\right] &\leq \mathcal{N}_{pd} \quad \text{(SF-1)} \\
\mathbb{E}\left[R_4^{14} + R_3^{1} + R_2^8 + R_1^{15}\right] &\leq \mathcal{N}_{pd} \quad \text{(SF-2)} \\
\mathbb{E}\left[R_4^{15} + R_3^{2} + R_2^9 + R_1^{16}\right] &\leq \mathcal{N}_{pd} \quad \text{(SF-3)} \\
\vdots \qquad & \qquad \vdots \\
\mathbb{E}\left[R_4^{12} + R_3^{19} + R_2^6 + R_1^{13}\right] &\leq \mathcal{N}_{pd} \quad \text{(SF-20)}
\end{aligned} \tag{3}$$

- ***Constraint 2:*** Since each subframe has a fixed number of PRBs, the number of initial (fresh) transmissions $\mathcal{F}^{\max}$ plus the expected number of re-transmissions (as discussed in ***Constraint 1***) should not exceed the total number of packets that can be filled in a subframe. Recalling that the maximum number of VoLTE packets that can be filled in a subframe is denoted as $\mathcal{N}_{vp}$, the second constraint for subframes $1-20$ can be expressed through the set of equations as shown below

in (4). :

$$\begin{aligned}
&\mathbb{E}\left[R_4^{13}+R_3^{20}+R_2^{7}+R_1^{14}\right]+\mathcal{F}^{\max}\leq \mathcal{N}_{vp} \quad \text{(SF-1)}\\
&\mathbb{E}\left[R_4^{14}+R_3^{1}+R_2^{8}+R_1^{15}\right]+\mathcal{F}^{\max}\leq \mathcal{N}_{vp} \quad \text{(SF-2)}\\
&\mathbb{E}\left[R_4^{15}+R_3^{2}+R_2^{9}+R_1^{16}\right]+\mathcal{F}^{\max}\leq \mathcal{N}_{vp} \quad \text{(SF-3)}\\
&\qquad\qquad\cdot\qquad\qquad\qquad\cdot\\
&\qquad\qquad\cdot\qquad\qquad\qquad\cdot\\
&\mathbb{E}\left[R_4^{12}+R_3^{19}+R_2^{6}+R_1^{13}\right]+\mathcal{F}^{\max}\leq \mathcal{N}_{vp} \quad \text{(SF-20)}
\end{aligned} \tag{4}$$

Note that $\mathcal{N}_{vp}$ is equal to $\lfloor \mathcal{N}_{prb}/\mathcal{M} \rfloor$, where $\lfloor . \rfloor$ is the floor function. Since the number of users who are allocated scheduling blocks for first transmission of their packets in each subframe is same and instantaneous received SNRs of users are independent, $\mathcal{F}^{\max}$ is the solution to the following optimization formulation:

$$\begin{aligned}
\max \quad & \mathcal{F}\\
\text{s.t.} \quad & \mathbb{E}\left[R_4+R_3+R_2+R_1\right]+\mathcal{F}\leq \mathcal{N}_{vp},\\
& \mathbb{E}\left[R_4+R_3+R_2+R_1\right]\leq \mathcal{N}_{pd},
\end{aligned} \tag{5}$$

where, the notation $\mathbb{E}[R_n]$ now denotes the expected number of packets from a subframe that require nth re-transmission (or $(n+1)$th transmission).

In order to solve this optimization problem, we will first find an expression for $\mathbb{E}[R_n]$. To do so, we define an indicator random variable X_k^n for user k's fresh packet in a subframe in the following manner.

$$X_k^n = \begin{cases} 1 & \text{if user } k\text{'s packet will require } n\text{th re-transmission}\\ 0 & \text{otherwise} \end{cases}$$

Since each subframe will have fresh packets for $\mathcal{F}$ users, we can now write,

$$\mathbb{E}[R_n] = \sum_{k=1}^{\mathcal{F}} P(X_k^n=1) = \mathcal{F}\times P(X_k^n=1), \tag{6}$$

in which the second equality is because users instantaneous received SNRs are independent. Hence, the optimization formulation (5) can be restated as:

$$\begin{aligned}
\max \quad & \mathcal{F}\\
\text{s.t.} \quad & \mathcal{F}\times\left(\sum_{n=1}^{4} P(X_k^n=1)+1\right)\leq \mathcal{N}\\
& \mathcal{F}\times\left(\sum_{n=1}^{4} P(X_k^n=1)\right)\leq \mathcal{N}_{pd}
\end{aligned} \tag{7}$$

Now, writing the first constraint as,

$$\mathcal{F}\leq \frac{\mathcal{N}_{vp}}{\sum_{n=1}^{4} P(X_k^n=1)+1}, \tag{8}$$

and the second constraint as,

$$\mathcal{F}\leq \frac{\mathcal{N}_{pd}}{\sum_{n=1}^{4} P(X_k^n=1)}. \tag{9}$$

We can replace the two constraints with a single constraint as,

$$\mathcal{F}\leq \min\left(\frac{\mathcal{N}_{vp}}{\sum_{n=1}^{4} P(X_k^n=1)+1}, \frac{\mathcal{N}_{pd}}{\sum_{n=1}^{4} P(X_k^n=1)}\right). \tag{10}$$

The $\mathcal{F}^{\max}$ that maximizes the objective in (5) can now be obtained by choosing an $\mathcal{F}$ that satisfies the equality in (10). Hence, we obtain $\mathcal{F}^{\max}$ as:

$$\mathcal{F}^{\max} = \min\left(\frac{\mathcal{N}_{vp}}{\sum_{n=1}^{4} P(X_k^n=1)+1}, \frac{\mathcal{N}_{pd}}{\sum_{n=1}^{4} P(X_k^n=1)}\right) \tag{11}$$

Hence, the capacity of E-PS-SPS is equal to

$$20\times\min\left(\frac{\mathcal{N}_{vp}}{\sum_{n=1}^{4} P(X_k^n=1)+1}, \frac{\mathcal{N}_{pd}}{\sum_{n=1}^{4} P(X_k^n=1)}\right) \text{ calls.} \tag{12}$$

In order for (12) to be useful, we need a closed form expression for $P(X_k^n=1)$ which we will derive next. Note that since X_k^n is an indicator random variable,

$$P(X_k^n=1) = 1-P(X_k^n=0). \tag{13}$$

For X_k^n to be 0, the effective instantaneous SNR on all PRBs within a user's scheduling block has to exceed the threshold SNR $\mathcal{S}_{th}$ after n transmissions. Since the instantaneous SNRs in each transmission attempt across PRBs within a scheduling block are i.i.d., the effective instantaneous SNRs across PRBs after n transmission attempts will also be i.i.d.. Hence,

$$P(X_k^n=0) = P\left(\bigcap_{i=1}^{\mathcal{M}} \gamma_{i,k,n}^{eff} > \mathcal{S}_{th}\right), \tag{14}$$

$$= \prod_{i=1}^{\mathcal{M}} P\left(\gamma_{i,k,n}^{eff} > \mathcal{S}_{th}\right), \tag{15}$$

$$= \left[P\left(\gamma_{k,n}^{eff} > \mathcal{S}_{th}\right)\right]^{\mathcal{M}}, \tag{16}$$

$$= \left[1-P\left(\gamma_{k,n}^{eff} \leq \mathcal{S}_{th}\right)\right]^{\mathcal{M}}. \tag{17}$$

Note that in the last two equalities above, $\gamma_{i,k,n}^{eff}$ is replaced with $\gamma_{k,n}^{eff}$ to signify that it is not dependent on PRB i within a user's scheduling block. Hence, we can now write,

$$P(X_k^n=1) = 1-\left[1-P\left(\gamma_{k,n}^{eff} \leq \mathcal{S}_{th}\right)\right]^{\mathcal{M}}. \tag{18}$$

Denoting the probability density function of the mean SNR μ_k of user k as $f_{\mu_k}(x)$ and using the total probability theorem, we can write,

$$P\left(\gamma_{k,n}^{eff} \leq \mathcal{S}_{th}\right) = \int_a^b P\left(\gamma_{k,n}^{eff} \leq \mathcal{S}_{th}/\mu_k = x\right) f_{\mu_k}(x)dx, \tag{19}$$

$$= \frac{1}{b-a}\int_a^b P\left(\gamma_{k,n}^{eff} \leq \mathcal{S}_{th}/\mu_k = x\right) dx, \tag{20}$$

since the mean SNR is distributed uniformly over $[a, b]$ for all users.

To deal with the definite integral of $P\left(\gamma_{k,n}^{eff} \leq \mathcal{S}_{th}/\mu_k = x\right)$, we first note that $\gamma_{k,n}^{eff}$ denotes the effective instantaneous received SNR on a PRB within user k's scheduling block after n transmissions of a VoLTE packet using chase combining for retransmissions. Now, given that the downlink instantaneous SNR on the PRB in each transmission is exponentially distributed with mean μ_k, using the results of Lemma 1 in [8], the probability density function of $\gamma_{k,n}^{eff}$ can be expressed as,

$$f_{\gamma_{k,n}^{eff}}(x) = \frac{x^{n-1}e^{-\frac{x}{\mu_k}}}{\Gamma(n)(\mu_k)^n}, \tag{21}$$

where, $\Gamma(n)$ is the gamma function, defined as, $\Gamma(n) = (n-1)!$.

Next, let r bits/RE be the rate[5] on each RE in every PRB of a scheduling block for transmitting VoLTE packet to user k experiencing a mean SNR μ_k. Let the downlink threshold SNR required for successful decoding of the data transmitted at rate r on each RE in a PRB be $\mathcal{S}_{th}$. Then, using the results of Lemma 2 in [8], the probability P_S^n of the data on any PRB being decoded successfully after n transmissions can be expressed as,

$$P_S^n = 1 - \frac{\gamma\left(n, \frac{\mathcal{S}_{th}}{\mu_k}\right)}{\Gamma(n)}, \tag{22}$$

where, $\gamma(a, b) = \int_0^b e^{-t}t^{a-1}\mathrm{d}t$ is the lower incomplete gamma function.

Now, noting that $P(\gamma_{k,n}^{eff} < \mathcal{S}_{th}/\mu_k = x) = 1 - P_S^n$, we can write,

$$P(\gamma_{k,n}^{eff} < \mathcal{S}_{th}/\mu_k = x) = \frac{\gamma\left(n, \frac{\mathcal{S}_{th}}{x}\right)}{\Gamma(n)}, \tag{23}$$

$$= \frac{\gamma\left(n, \frac{\mathcal{S}_{th}}{x}\right)}{(n-1)!}. \tag{24}$$

Hence, (19) can now be written as,

$$P\left(\gamma_{k,n}^{eff} \leq \mathcal{S}_{th}\right) = \frac{1}{b-a}\int_a^b \frac{\gamma\left(n, \frac{\mathcal{S}_{th}}{x}\right)}{\Gamma(n)}dx. \tag{25}$$

In order to integrate the lower incomplete gamma function, we write it as,

$$\gamma\left(n, \frac{\mathcal{S}_{th}}{x}\right) = (n-1)!\left[1 - e^{-\frac{\mathcal{S}_{th}}{x}}\sum_{k=0}^{n-1}\frac{(\mathcal{S}_{th}/x)^k}{k!}\right]. \tag{26}$$

Hence,

$$P\left(\gamma_{k,n}^{eff} \leq \mathcal{S}_{th}\right)$$

$$= \frac{1}{(b-a)}\int_a^b \left[1 - e^{-\frac{\mathcal{S}_{th}}{x}}\sum_{k=0}^{n-1}\frac{(\mathcal{S}_{th}/x)^k}{k!}\right]dx, \tag{27}$$

$$= \frac{1}{(b-a)}\left[(b-a) - \sum_{k=0}^{n-1}\frac{\mathcal{S}_{th}^k}{k!}\int_a^b e^{-\frac{\mathcal{S}_{th}}{x}}x^{-k}\right]dx. \tag{28}$$

[5]Let $\mathcal{M}$ be the number of PRBs in a scheduling block and let D be the number of data carrying REs in each PRB. Then if the size of the VoLTE packet is $\mathcal{L}$ bits, the rate $r = \mathcal{L}/\mathcal{M}D$ bits/RE.

Simplification of the inner integral yields,

$$P\left(\gamma_{k,n}^{eff} \leq \mathcal{S}_{th}\right) =$$

$$1 - \frac{1}{b-a}\sum_{k=0}^{n-1}\frac{\mathcal{S}_{th}}{k!}\left[\Gamma\left(k-1, \frac{\mathcal{S}_{th}}{b}\right) - \Gamma\left(k-1, \frac{\mathcal{S}_{th}}{a}\right)\right], \tag{29}$$

which can be plugged into (18) to obtain $P(X_k^n = 1)$ that is required to get the capacity of E-PS-SPS in (12).

C. Success Rate

E-PS-SPS facilitates maximum possible number of transmissions of all packets of all users till the capacity is reached. However, the effective received SNR $\gamma_{k,n}^{eff}$ for a fraction of packets of a user k may not reach the threshold SNR $\mathcal{S}_{th}$ on all PRBs of the scheduling block even after 5 transmissions due to inadequate mean SNR. This leads to the call being termed unsatisfactory when the fraction of failed packets of the user exceeds a certain threshold. Hence, there is a maximum success rate that can be achieved since all the connected users may not have the adequate mean SNR required for the satisfaction of their call. Let N_C be the number of users that got connected and N_S be the number of users satisfied with their calls. We define the success rate S_R as,

$$S_R = \lim_{N_C\to\infty}\frac{N_S}{N_C} \times 100\%. \tag{30}$$

A user is said to be satisfied with a call if the user satisfaction ratio U_R exceeds a threshold α, i.e.,

$$U_R = \lim_{N_A\to\infty}\frac{N_R}{N_A} \geq \alpha, \tag{31}$$

where, N_A is the total number of packets that arrived for a user, N_R is the number of packets that reached the user correctly and α is the threshold for the call to be satisfactory.

Hence, the success rate S_R is simply the fraction of the total connected devices for whom the success probability of each packet exceeds α. Denoting the minimum downlink mean SNR required for the success probability of a packet to exceed α as μ_{min} and with the mean SNRs of the connected users distributed uniformly over $[a, b]$, we can write S_R as,

$$S_R = \frac{b - \mu_{\min}}{b-a} \times 100\%. \tag{32}$$

where, using (23), the minimum mean SNR $\mu_{\min}$ can be obtained as:

$$\mu_{\min} = \arg\max_x\left[1 - \frac{\gamma\left(5, \frac{\mathcal{S}_{th}}{x}\right)}{(5-1)!}\right]^{\mathcal{M}} < \alpha \tag{33}$$

Observe that once the number of calls exceeds the capacity, the success rate falls below S_R since E-PS-SPS fails to facilitate maximum possible number of packet transmissions resulting in packet failures even for users with adequate mean SNRs. Hence, the capacity of E-PS-SPS can also be defined as the number of calls beyond which the success rate falls below S_R.

Table 2. Table showing the summary of results obtained using the derived analytical expressions and Monte Carlo simulations.

No. of PRBs $\mathcal{M}$ for each VoLTE packet.	Rate r in bits per resource element.	Required threshold SNR $\mathcal{S}_{th}$ in dB	Required min. mean SNR μ_{min} in dB for $\alpha = 0.98$ from (33).	Success rate S_R for $\alpha = 0.98$ from (32).	Success rate S_R for $\alpha = 0.98$ from simulations for E-PS-SPS	Required min. mean SNR μ_{min} in dB for $\alpha = 0.95$ from (33).	Success rate S_R for $\alpha = 0.95$ from (32)	Success rate S_R for $\alpha = 0.95$ from simulations for E-PS-SPS	Capacity $\mathcal{C}$ for 10 MHz bandwidth (50 PRBs) from (29) and (12).	Capacity $\mathcal{C}$ for 10 MHz bandwidth (50 PRBs) from simulations for E-PS-SPS
1	2.5	10.68	6.98	65.5%	67%	5.42	73%	77%	52 calls	55 calls
2	1.25	5.39	4.20	79.0%	79%	3.3	83.5%	85%	83 calls	80 calls
3	0.83	2.93	2.52	87.5%	84%	2.0	90%	89%	130 calls	125 calls
4	0.62	1.34	1.24	93.8%	91%	0.98	95.1%	94%	180 calls	176 calls
5	0.50	0.17	0.16	99.2%	98.5%	0.13	99.35%	99%	188 calls	185 calls
6	0.41	-0.75	0.0	100%	100%	0.0	100%	100%	160 calls	160 calls

IV. ANALYSIS AND SIMULATION RESULTS

In this section, we will present the numerical results obtained from the analytical expressions and compare them with results obtained from Monte Carlo simulations.

A. Numerical Results using Derived Analytical Expressions

We consider a cell of 10 MHz bandwidth. Hence, the number of PRBs in each subframe $\mathcal{N}_{prb} = 50$. We take $\mathcal{N}_{pd} = 4$ as done in [8]. We assume the cell edge user's mean SNR to be 0 dB and the cell center user's mean SNR to be about 20 dB [24]. Hence, each connected user's downlink mean SNR is a uniform random variable over $[0, 20]$ dB interval. Hence $a = 0$ and $b = 20$. VoLTE packet size $\mathcal{L}$ is set to 300 bits [8], [20], and the number of data carrying REs D within a PRB be 120 [8]. We will obtain the capacity and success rate values for the scheduling block size $\mathcal{M}$ (ranging from 1 to 6 PRBs) used to transmit a VoLTE packet. For demonstration, we show here the procedure to use the various analytical expressions to find the success rate and the capacity for $\mathcal{M} = 1$ and $\mathcal{M} = 3$. A similar procedure is applied for other values of $\mathcal{M}$.

For $\mathcal{M} = 1$, i.e., 1 PRB, we have a total of 120 REs for 300 bits. Hence, the rate r required to carry 300 bits in 120 REs would be 2.5 bits per RE. Also, $\mathcal{N}_{vp}$, i.e., the number of VoLTE packets that can be accommodated is equal to 50 for $\mathcal{M} = 1$.

We consider the following relationship between the rate r and threshold SNR $\mathcal{S}_{th}$ [8], [26]:

$$r = \log_2 \left(1 + \eta \mathcal{S}_{th}\right), \tag{34}$$

where, η is the coding gain loss. We use $\eta = 0.398$ as done in [26]. A packet is then successfully decoded after the nth transmission to a user k if $\gamma_{k,n}^{eff} \geq \mathcal{S}_{th}$ for all PRBs within the scheduling block, where $\mathcal{S}_{th}$ satisfies,

$$0.8330 = \log_2 \left(1 + 0.398 \mathcal{S}_{th}\right).$$

Solving the above equation yields a value of 10.68 dB for $\mathcal{S}_{th}$. We then have from our derived eqn. (29), $P(X_i^1 = 1) = 0.69$, $P(X_i^2 = 1) = 0.41$, $P(X_i^3 = 1) = 0.25$, and $P(X_i^4 = 1) = 0.17$. Plugging these values into the derived eqn. (12), we get a capacity of 52 calls.

Similarly, for $\mathcal{M} = 3$, we have a total of 360 resource elements. Hence, the rate r would be 0.83 bits per resource element yielding a value of 2.93 dB for $\mathcal{S}_{th}$. Hence, we get $P(X_i^1 = 1) = 0.35$, $P(X_i^2 = 1) = 0.13$, $P(X_i^3 = 1) = 0.07$, and $P(X_i^4 = 1) = 0.04$. Plugging these values into (12) gives a capacity of 130 calls.

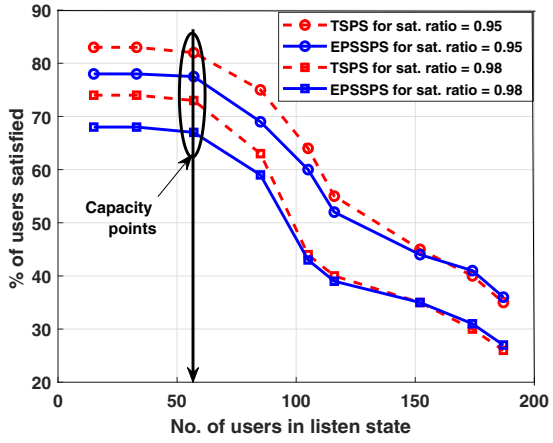
We also calculate the success rate S_R for a user satisfaction ratio of $\alpha = 0.98$ (i.e., at least 98% of the user's downlink VoLTE packets should reach successfully for the user to be satisfied with the call) and $\alpha = 0.95$. For $\alpha = 0.98$ and $\mathcal{M} = 1$, using (33), we first obtain $\mu_{\min} = 6.98$ dB. Plugging this value into (32), we get $S_R = 65.50\%$. For $\mathcal{M} = 3$, we obtain $\mu_{\min} = 2.52$ dB giving a success rate $S_R = 87.50\%$. Success rate for other values of $\mathcal{M}$ and α can be similarly obtained. A summary of the values thus obtained using analytical expressions is provided in Table 2. The table also contains results obtained from Monte Carlo simulations as discussed next.

B. Simulation Procedure

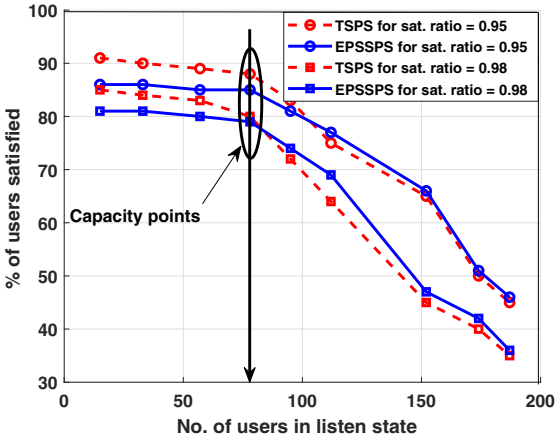
In order to validate the above capacity and success rate values that are obtained using our derived analytical expressions, we next carry out Monte Carlo simulations. We implement the simulation code in C programming language. For simulations, along with the parameter values already mentioned above, the instantaneous SNR in each transmission of a VoLTE packet is modeled as an exponentially distributed random variable. The effective instantaneous SNR in the nth transmission is the sum of individual instantaneous SNRs of each transmission attempt.

We simulate the system for 10 hours operation, i.e., 36000000 subframes for it to reach and be in steady state for sufficient time duration. We implement the call arrivals as a Poisson process with the inter-arrival times being exponentially distributed. Each call duration is taken as an exponential random variable with a mean of 120 s. Within each call, a user alternates between talking and listening states according to exponential distribution with a mean of 4 s each. VoLTE packets for a user arrive every 20 ms when in listening state.

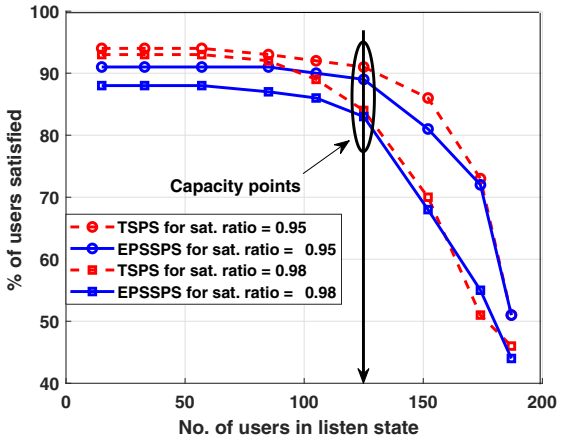
The operation of E-PS-SPS is as described in Section II.D. In every subframe, E-PS-SPS scheduler starts by closing the calls whose time duration is completed and releasing the PRBs that were allocated to that user. It then schedules the fresh packets within that subframe according to E-PS-SPS schedule. The previously failed packets waiting in HARQ queues are then taken up for re-transmission till the free PRBs and/or PDCCH signals are exhausted. After HARQ packets, the scheduler handles users going into talking mode by releasing downlink PRBs that were



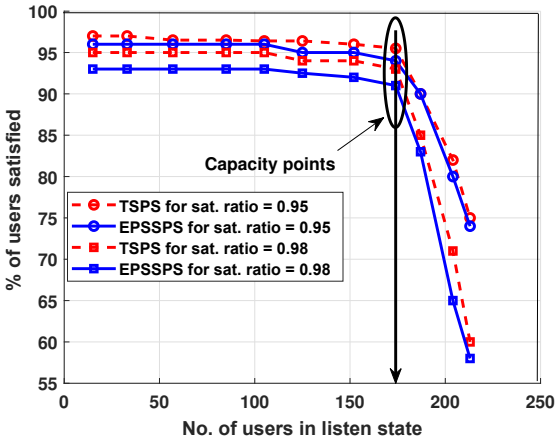
(a)



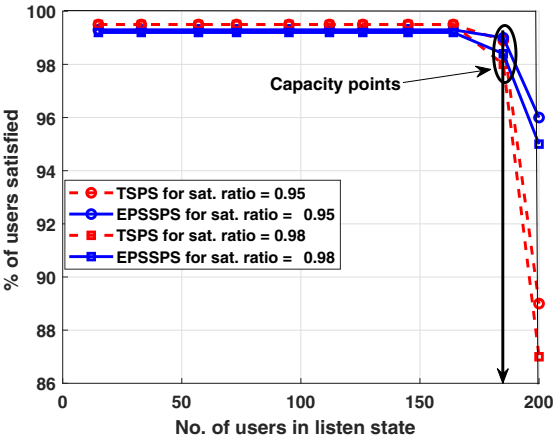
(b)



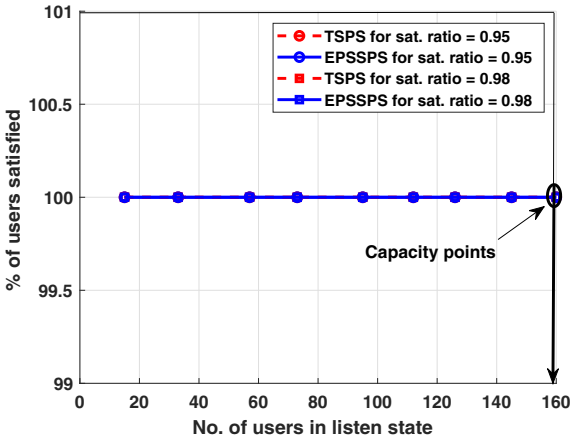
(c)



(d)



(e)



(f)

Fig. 3. Each plot shows simulation results of % success rate (percentage of users satisfied) V/S number of users in listening state for satisfaction ratios of $\alpha = 0.95$ and $\alpha = 0.98$ with T-SPS and E-PS-SPS in downlink operation, for a 10 MHz bandwidth (50 PRBs) cell. The number of users beyond which the success rate starts degrading is the capacity point as shown in each of the plots. We show the results for number of PRBs used per VoLTE packet varying from 1 (Fig. 3(a)) to 6 (Fig. 3(f)): (a) Scheduling block size = 1 PRB, (b) scheduling bock size = 2 PRBs, (c) scheduling bock size = 3 PRBs, (d) scheduling block size = 4 PRBs, (e) scheduling block size = 5 PRBs, and (f) scheduling block size = 6 PRBs.

allocated during the listening mode and finally it handles any fresh calls arriving in that subframe.

We simulate the behavior of the system for the downlink operation of E-PS-SPS. We run the simulation for a duration of 36000000 subframes with a fixed arrival rate of calls. For each arrival rate we calculate the success rate. We repeat this simulation for each arrival rate from 0.1 calls/s in steps of 0.2 till the success rate starts degrading. The user satisfaction ratio U_R for a connected user is calculated as the ratio of number of packets that reached the user correctly within the delay constraint of 50 ms to the total number of packets that arrived for this user at the eNodeB. The user is considered satisfied if $U_R > \alpha$ as assumed in the analysis procedure. The success rate S_R is then calculated for each arrival rate as the ratio of number of users satisfied (i.e., $U_R > \alpha$) to the total number of users connected within that simulation round. We simulate the system for scheduling block sizes from $\mathcal{M} = 1$ PRB to $\mathcal{M} = 6$ PRBs in order to compare the results with the ones obtained using analytical expressions above. For each value of $\mathcal{M}$, we calculate the success rate for each arrival rate of calls and also count the maximum number of users that were in listening state simultaneously at any point of time during simulation for that arrival rate. We then plot the success rates versus the maximum number of calls in listening state for each arrival rate. The capacity is then obtained as the point beyond which the success rate starts degrading. We find the capacity for $\alpha = 0.95$ and $\alpha = 0.98$.

C. Observations and Discussions

Fig. 3 shows the results obtained from simulations. Consider the success rate plots in Fig. 3(a). It can be seen that the success rate for E-PS-SPS for a satisfaction ratio of 0.98 is 67% till the number of users in listen state reaches about 55. Hence, this point of 55 users beyond which the success rate starts degrading is taken as the capacity for a scheduling block size of 1 PRB. Some observations and insights are in order. For a scheduling block size of 1, note that a total of 50 VoLTE packets can be accommodated in a single subframe of 50 PRBs making it a total of 1000 fresh VoLTE packets (users) in a cycle of 20 subframes. However, what Fig. 3(a) shows is that with only 4 PDCCHs available in each subframe to support re-transmissions, once the number of fresh VoLTE packets in a subframe crosses 3, the average number of packets that will require re-transmission in a subframe crosses 4. Hence with only 4 PDCCHs available in a subframe, the system cannot ensure enough number of re-transmissions for a packet within its delay constraint (4 re-transmissions within 50 ms delay constraint) for it to reach the user with the minimum required effective SNR. Therefore, as seen from Fig. 3(a), beyond the point of 55 users in listen state (i.e., roughly about 3 fresh packets per subframe), the success rate starts degrading as users fail to get satisfied with increasing number of packet failures. Hence, the capacity for a scheduling block size of 1 PRB is taken as 55 calls.

When the scheduling block size is increased beyond 1 PRB, the rate of transmission r decreases and so does the required threshold SNR $\mathcal{S}_{th}$. Hence, the number of re-transmissions also decrease allowing for more fresh packets in each subframe. Due to this, the point beyond which the success rate starts degrading keeps getting pushed to the right for increasing scheduling block size. However, there is still a point for each scheduling block size beyond which the number of PDCCHs start falling short to support the increasing re-transmissions. Hence, till the scheduling block size of 5 PRBs, the capacity keeps increasing (Fig. 3(a) to Fig. 3(e)) but remains PDCCH limited. However, for a scheduling block size of 6 PRBs, the required threshold SNR becomes so low that all the packets are successful in their first transmission itself. Hence, PDCCHs are not needed at all because of the absence of re-transmissions. However, with a scheduling size of 6 PRBs, a total of only 8 fresh packets can be accommodated in a subframe giving a capacity of 160 calls as seen in Fig. 3(f). Hence, beyond a scheduling block size of 5, the system capacity starts decreasing and is PRB limited. The results of the simulations are also summarized in Table 2 to compare with the results obtained from the derived analytical expressions. It can be seen that the simulation results match very closely with analysis results thus validating our derived expressions.

D. Optimal Operation of E-PS-SPS

Having validated our analytical expressions for the capacity and success rate of E-PS-SPS, we now demonstrate how E-PS-SPS can be operated optimally for various parameters. In order to do so, we use the results from Table 2. One can now choose the optimal size of a scheduling block based on the success rate requirement. For example, if the desired success rate is 99% for a user satisfaction ratio $\alpha = 0.98$ for a 10 MHz cell (50 PRBs in each subframe), then from the table, it can be seen that using a scheduling block of size 5 PRBs is optimal since it provides a capacity of 188 calls. However, if the desired success rate is say 100%, then using scheduling blocks of 6 PRBs is optimal giving a capacity of 160 calls. As already explained, using more than 6 PRBs also offers 100% success rate, but the capacity will be lesser than 160 since we can only accommodate less than 8 packets in a subframe with 50 PRBs, for a scheduling block size of more than 6 PRBs. Therefore a scheduling block of size more than 6 PRBs is sub-optimal for a 100% success rate for a 10 MHz cell.

Further, it can be noted from (32) and (33) that the success rate only depends on user satisfaction ratio threshold and the required threshold SNR which in turn depends on size of the scheduling block (number of PRBs used to transmit the VoLTE packet) only. Hence, the success rate is independent of the bandwidth of the system. However, the number of PDCCH signals and the number of VoLTE packets in a subframe scale linearly with bandwidth. For instance, if the cell bandwidth is doubled (say from 10 MHz to 20 MHz), then the number of PDCCH signals also double from 4 to 8 and the number of VoLTE packets in a subframe also double for a fixed scheduling block size. Hence, from (12) and (29) it can be seen that the capacity of E-PS-SPS also scales linearly with bandwidth by the same factor. Therefore, the optimal performance results for different system parameters like user satisfaction ratio threshold α and bandwidth of the cell can be easily obtained from our derived analytical expressions.

E. Comparison with T-SPS

We also carry out simulations for T-SPS to compare its performance with E-PS-SPS. As seen from the plots in Fig. 3 in all cases, the capacity of T-SPS is same as E-PS-SPS. However, T-SPS shows a marginal improvement in the success rate over E-PS-SPS. This is because a fresh packet gets scheduled by T-SPS in the first subframe that has free PRBs after its arrival. But E-PS-SPS schedules according to a sequence that ensures energy saving. This results in lesser initial average delay for the packet under T-SPS scheduling that helps it to have one extra re-transmission within 50 ms delay constraint after repeated failures. This improves the chances of the packet reaching the user with an effective SNR greater than the threshold SNR thus increasing the success rate. However, this increase in success rate is very marginal and hence can be overlooked in favor of significant energy saving that E-PS-SPS can provide as shown in [8] when the system is operating below the capacity point.

V. CONCLUSION

In this paper, we presented an accurate performance analysis of E-PS-SPS by deriving closed form expressions for its capacity and success rate in a cell with VoLTE users having i.i.d. mean SNRs and i.i.d. instantaneous SNRs across PRBs within a user's scheduling block. After validating with Monte Carlo simulations, we desribed how these expressions can be used to find the optimal size of the scheduling block in order to maximize the capacity of E-PS-SPS, given a desired success rate. For example, for a 10 MHz cell, our work shows that the optimal number of PRBs to be allocated to each VoLTE user is 6 to achieve a maximum capacity of 160 VoLTE calls if a 100% satisfaction rate is desired. Our results also show that the capacity can be increased upto 188 calls which is possible if the satisfaction rate is reduced to 99%. Another interesting insight obtained from our work is that the capacity cannot be increased beyond 188 calls even if the desired satisfaction rate is reduced below 99%. Similar results can be easily obtained for cells for different bandwidths and user satisfaction ratios using our analysis. We also showed through extensive simulations that the capacity of E-PS-SPS is same as T-SPS in a general setting which we considered in the work. Further, our work also revealed that T-SPS has a marginally better success rate than E-PS-SPS. However, this marginal loss in success rate can be overlooked considering the significant energy saving that E-PS-SPS can offer.

REFERENCES

[1] Z. Hasan, H. Boostanimehr, and V. K. Bhargava, "Green cellular networks, a survey, some research issues and challenges", *IEEE Commun. Surveys Tuts.*, vol. 13, no. 4, pp. 524–540, Nov. 2011.

[2] P. Frenger *et al.*, "Reducing energy consumption in LTE with cell DTX", in *Proc. IEEE VTC*, 2011.

[3] J. Lorincz, T. Garma, and G. Petrovic, "Measurements and modeling of base station power consumption under real traffic loads", in *Sensors*, vol. 12, pp. 4281–4310, 2012.

[4] A. Conte *et al.*, "Cell wilting and blossoming for energy efficiency", *IEEE Wireless Commun.*, vol. 18, no. 5, pp. 50–57, Oct. 2011.

[5] R. Gupta and E. C. Strinati, "Base-station duty-cycling and traffic buffering as a means to achieve green communications", in *Proc. IEEE VTC*, 2012.

[6] L. Saker, S. Elayoubi, and T. Chahed, "Minimizing energy consumption via sleep mode in green base station," in *Proc. IEEE WCNC*, 2010.

[7] P. K. Wali and D. Das, "PS-SPS: Power saving-semi persistent scheduler for VoLTE in LTE-advanced", in *Proc. IEEE CONECCT*, 2015.

[8] P. K. Wali and D. Das, "Enhanced-power saving semi-persistent scheduler for VoLTE in LTE-advanced," *IEEE Trans. Wireless Commun.*, vol. 15, no. 11, pp. 7858–7871, Nov. 2016.

[9] D. P. Basavaraj *et al.*, "Throughput and service enhancing network selection algorithm for dual registration in 5G standalone architecture," in *Proc. IEEE ANTS*, 2019.

[10] Y. Feng, A. Nirmalathas, and E. Wong, "A predictive semi-persistent scheduling scheme for low-latency applications in LTE and NR networks," in *Proc. IEEE ICC*, 2019.

[11] N. Afrin, J. Brown, and J. Y. Khan, "Design of a buffer and channel adaptive LTE semi-persistent scheduler for M2M communications," in *Proc. IEEE ICC*, 2015.

[12] S. Kim, W. Cha, J. So, M. Na, and C. Choi, "Performance evaluation of semi-persistent scheduling in a narrowband LTE system for Internet of things," *J, Korean Institute Commun. Inf. Sci.*, vol. 41, no. 9, pp. 1001-–1009, Sept. 2016.

[13] A. Nabil, K. Kaur, C. Dietrich, and V. Marojevic, "Performance analysis of sensing-based semi-persistent scheduling in C-V2X networks," in *Proc. IEEE VTC*, 2018.

[14] A. González, N. Franchi, and G. Fettweis, "Control loop aware LTE-V2X semi-persistent scheduling for string stable CACC," in *Proc. IEEE PIMRC*, 2019.

[15] Z. Arnjad, A. Sikora, B. Hilt, and J. Lauffenburger, "Latency reduction for narrowband LTE with semi-persistent scheduling," in *Proc. IEEE IDAACS-SWS*, 2018.

[16] C. Pan, Y. Yang, and Y. Li, "A priority based semi-persistent scheduling for VoLTE," in *Proc. IEEE ICIEA*, 2018.

[17] S. Cho, Y. Kim, and Y. Han, "Analysis of control overhead reduction by semi-persistent scheduling," *IEEE Wireless Comm. Lett.*, vol. 9, no. 5, pp. 726–730, May 2020.

[18] Farooq Khan, *LTE for 4G mobile broadband air interface technologies and performance*. Richardson, TX: Cambridge University Press, 2009.

[19] S. Sesia, I. Toufik, and M. Baker, *LTE – the UMTS long term evolution, from theory to practice*, John Wiley and Sons, 2009.

[20] D. Jiang *et al*, "Principle and performance of semi-persistent scheduling for VoIP in LTE system ", in *Proc. WiCom*, 2007.

[21] H. Holma and A. Toskala, *LTE for UMTS-OFDMA and SC-FDMA based radio access*, Wiley; 2 edition (April 25, 2011).

[22] S. N. Donthi and N. B. Mehta, "Performance analysis of subband-level channel quality indicator feedback scheme of LTE", in *Proc. NCC*, 2010.

[23] John Proakis, *Digital Communications*. 4th ed. New York: McGraw-Hill, 2000.

[24] J. Francis and N. B. Mehta, "Throughput-optimal scheduling and rate adaptation for reduced feedback best- M scheme in OFDM systems," *IEEE Trans. Commun.*, vol. 65, no. 7, pp. 3053–3065, July 2017.

[25] D. Parruca and J. Gross, "Rate selection analysis under semi-persistent scheduling in LTE networks," in *Proc. ICNC*, 2013.

[26] S. N. Donthi and N. B. Mehta, "Performance analysis of user selected subband channel quality indicator feedback scheme of LTE", in *Proc. IEEE GLOBECOM*, 2010.

[27] G. Piro, L. A. Grieco, G. Boggia, R. Fortuna, and P. Camarda, "Two-level downlink scheduling for real-time multimedia services in LTE networks", *IEEE Trans. Multimedia*, vol. 13, no. 5, pp. 1052–1065, Oct. 2011.

Prashant Wali is currently a Faculty in the department of Electrical and Electronics Engineering at BITS Pilani, Hyderabad campus. He received his BE degree in Electronics and Communication from PDA College of Engineering, Gulbarga in 2000. He received his Masters degree from the Dept.of Electrical Communication Engineering, Indian Institute of Science, Bangalore, India in 2008 and PhD from IIIT-Bangalore, India in 2017. His research interests include design and analysis of energy efficient algorithms for broadband wireless cellular systems.