Authors' version of:
G. Stea, A. Virdis, "A comprehensive simulation analysis of LTE Discontinuous Reception (DRX)", Elsevier Computer Networks, 73 (2014), pp.22-40, DOI 10.1016/j.comnet.2014.07.014

A comprehensive simulation analysis of LTE Discontinuous Reception (DRX)

Giovanni Stea, Antonio Virdis

***Abstract*— In an LTE cell, Discontinuous Reception (DRX) allows the central base station to configure User Equipments for periodic wake/sleep cycles, so as to save energy. DRX operations depend on several parameters, which can be tuned to achieve optimal performance with different traffic profiles (i.e., CBR vs. bursty, periodic vs. sporadic, etc.). This work investigates how to configure these parameters and explores the trade-off between power saving, on one side, and per-user QoS, on the other. Unlike previous work, chiefly based on analytical models neglecting key aspects of LTE, our evaluation is carried out via simulation. We use a fully-fledged packet simulator, which includes models of all the protocol stack, the applications and the relevant QoS metrics, and employ factorial analysis to assess the impact of the many simulation factors in a statistically rigorous way. This allows us to analyze a wider spectrum of scenarios, assessing the interplay of the LTE mechanisms and DRX, and to derive configuration guidelines.**

***Index Terms*—LTE, DRX, Resource Allocation, Quality of Service, Power Saving, Simulation**

I. Introduction

THE Long-Term Evolution (LTE) of the UMTS promises ubiquitous, high-speed Internet access. In such systems, a central base station or enhanced-NodeB (eNB) shares radio resources among a number of User Equipments (UEs), i.e. handheld devices, laptops or home gateways. Handheld devices are normally battery-powered, hence care must be taken not to waste energy. On the network side, this objective can be aided by properly configuring *Discontinuous Reception* (DRX), which allows UEs to power off the reception/transmission circuitry periodically, waking up for short periods at specific instants. The underlying rationale is that packet transmission/reception is hardly ever continuous over time, hence synchronizing it with wake-up periods is likely to achieve significant energy savings with only a moderate increase in latency. The UE DRX is configured by the eNB semi-statically, by tuning several parameters: the *cycle length*, the *on* duration and offset within the cycle; the *inactivity timer*, which prolongs the *on* duration when a packet arrives, thus coping with bursty arrivals; the *short vs. long cycle*, which allows an UE to power down for several short intervals and check for new packets before going to sleep for longer times. These parameters can only be varied with a signaling procedure that takes hundreds of milliseconds, hence cannot follow short-term traffic variations. A more dynamic feature of DRX is instead the *sleep* control message, by which the eNB can send UEs to sleep until their next scheduled wake-up time.

Giovanni Stea and Antonio Virdis are with the Dipartimento di Ingegneria dell'Informazione of the University of Pisa, Largo Lucio Lazzarino 1, 56122 Pisa, Italy (e-mail: g.stea@iet.unipi.it, a.virdis@iet.unipi.it).

A large number of papers have recently evaluated the performance of DRX under various conditions ([6]-[29]). Most of these studies define analytical models to capture the essential behavior of DRX-enabled UEs with different types of traffic. Our experience is that LTE modeling is a complex task, since it involves a considerable amount of submodels, themselves often complex enough as to defy analytical modeling, and interacting with each other in complex ways: physical channel, MAC protocol with fragmentation and H-ARQ, resource allocation, application behavior, time- and location-varying channel quality, etc. To the best of our knowledge, none of the above works compare the results obtained with their analytical models with those obtained in a setting that models the above features. Some works that analyze the DRX performance via simulation have appeared recently (e.g., [23]). Simulation-based investigation lends itself to more detailed modeling. However, these works study a limited number of scenarios and traffics (typically only the downlink and VoIP), and neglect some features that instead play a crucial role in DRX performance.

Our claim is that the DRX performance, and - specifically – the trade-off between QoS and power consumption, depends on a multitude of factors: the traffic profile and requirements, the cell load, the access methods employed at the eNB, and – of course – the manifold DRX settings. To gain insight into this, a systematic approach is required.

In this work, we analyze the performance of DRX, with the aim to obtain configuration guidelines and estimates of its impact on the whole cell. We carry out this study via simulation, using a fully-fledged C++ simulator which includes detailed models of all the layers and functions of LTE, models of applications and mobility, and relevant QoS and Quality of Experience (QoE) metrics. We employ *factorial analysis* [27] to determine the impact of the parameters on the relevant metrics in a statistically rigorous way. We study DRX configuration for several applications: symmetric (VoIP), asymmetric (HTTP web browsing and YouTube video) and downlink-only (streaming Video on Demand). Our results show that the trade-off between power consumption and QoS is generally favorable, meaning that a considerable power reduction is achieved by giving in a tolerable QoS degradation. Moreover, the tradeoff can be fine-tuned: when the load increases, less aggressive DRX settings can be used to safeguard performance.

The rest of the paper is organized as follows: in Section II we provide the reader with the necessary background on the LTE and DRX standards. We describe our simulation methodology, tools and settings in Section III, and report performance evaluation results in Section IV. Section V reviews the related

Authors' version of:
G. Stea, A. Virdis, "A comprehensive simulation analysis of LTE Discontinuous Reception (DRX)", Elsevier Computer Networks, 73 (2014), pp.22-40, DOI 10.1016/j.comnet.2014.07.014

work. Finally, Section VI reports conclusive remarks.

II. Background on LTE

Hereafter we describe the aspects of the LTE system which are more relevant to the resource allocation problem in both the downlink and uplink directions. A table of LTE-related acronyms is reported in the Appendix for ease of reference.

In LTE, transmissions are arranged in frames, called Transmission Time Intervals, (TTIs), whose duration is 1ms. In the downlink, the eNB allocates a vector of *Resource Blocks* (RBs) to the UEs associated to it on each TTI, by broadcasting the RB allocation map in the Physical Downlink Control Channel (PDCCH) (see Figure 1). Each RB carries a fixed number of symbols, which translate to different amounts of bits depending on the modulation and coding scheme used by the UE. In general, UEs favor more information-dense modulations (e.g., up to 64QAM, which yields 6 bits per symbol) when they perceive a better channel to the eNB. The quality of the wireless channel is time-varying, hence UEs report their perceived channel state to the eNB as a Channel Quality Indicator (CQI), periodically (e.g., every 5 ms) or on demand. The latter is an index in a standard table, computed by the UE according to the measured Signal to Interference and Noise Ratio (SINR), and determines the modulation that the latter will use. The amount of information being sent to a UE in a TTI, encapsulated in a Protocol Data Unit (PDU) is called Transmission Block Size (TBS). An exemplary mapping is reported in Table 1, with the caveat that the number of bytes transmitted in a RB is not a constant function of the CQI, but also depends on the number of RBs on which the TBS is coded. Transmissions are subject to errors, and are therefore protected by a Hybrid ARQ (H-ARQ) scheme, which allows a configurable number of retransmissions. Downlink H-ARQ processes are *asynchronous*, meaning that they are part of the eNB scheduling: a given retransmission may take place at any future TTI, when the eNB schedules the relevant H-ARQ process.

In the uplink, the UE notifies the eNB about its backlog state via quantized *Buffer Status Reports* (*BSRs*). BSRs are transmitted (either alone or trailing a data transmission) *in-band*, i.e. together with the data. Thus, they can only be sent i) when the UE is scheduled, and ii) if there is enough space to do so (a BSR can take up to 24 bits). Therefore, a mechanism is needed to allow a UE to signal its transition from empty to backlogged. UEs signal their service requests *out of band*, using a dedicated Random Access Procedure (RAC) and a backoff mechanism to arbitrate collisions. RAC requests are instead responded in-band, by scheduling the UE in a future TTI[1]. RAC requests are re-iterated after a random period of time if the UE is not scheduled. The standard handshake for uplink transmissions, shown in Figure 2, takes five messages: first the UE initiates a RAC request; then, the eNB responds by issuing a short grant, large enough for a BSR; the UE sends its BSR; the eNB sends a larger grant according to some scheduling policy, and finally the UE transmits its data. In some cases (e.g., when uplink traffic is predictable), the eNB may decide to dispense with the middle two interactions, and immediately issue a grant large enough to hold the BSR *and* one or more PDUs in response to the RAC request. This technique, called *bandwidth stealing*, is known to increase the uplink capacity and reduce the latency.

Semi-Persistent Scheduling (SPS, [4]) can also be used for uplink transmissions of periodic, low-bandwidth traffic, e.g., VoIP. It consists in the eNB issuing periodic grants to the UEs, which can then transmit without the need for signaling or handshake in the pre-assigned TTIs. A periodic grant can be revoked *explicitly*, via a specific message, or *implicitly*, after the UE fails to exploit it for a given number of consecutive times. Note that, under SPS, the periodic grant also sets – once and for all – the *format* of the uplink transmission, thus preventing link adaptation. Hence, variations in the channel quality (which are unavoidable, especially in the long term) may increase the Block Error Rate (BLER) or force the eNB to overdimension the periodic grant, thus reducing the efficiency of the scheduling process. Uplink H-ARQ processes are *synchronous*, i.e., they alternate over a period of eight TTIs. This means that an uplink retransmission takes place exactly after eight TTI have elapsed from the previous one.

Finally, we observe that the eNB participates in flow signaling, hence is able to classify flows. The type of flow can be encoded in the QoS Class Identifier (QCI), e.g. QCI 1 for conversational voice, QCI 7 for live video streaming, etc.

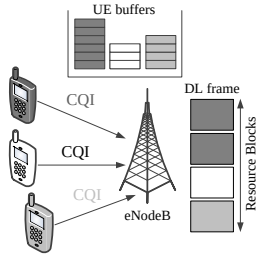

Figure 1 – Scheduling of downlink connections in LTE

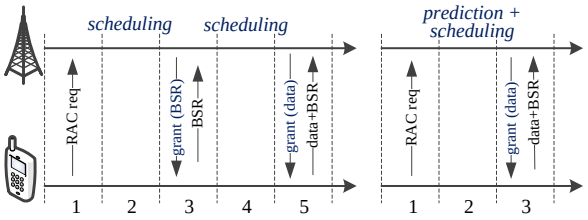

Figure 2 – Handshake for scheduling of uplink UE traffic: standard (left) and using Bandwidth Stealing (right).

[1] The standard also defines a *Dedicated Scheduling Request (DSR)* mode, whereby UEs issue scheduling requests using in-band *dedicated* resources. DSR is increasingly inefficient as the number of UEs grows large, hence it is scarcely used in practice and will not be considered further in this work.

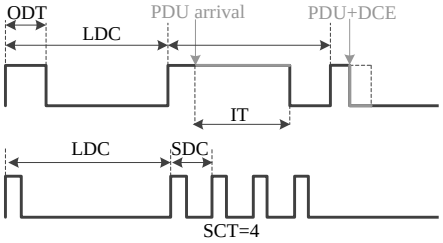

Figure 3 – Basic mechanisms for DRX; inactivity timer and DCE (top) and long/short cycles (bottom).

TABLE 1 – EXEMPLARY CQI MAPPING.

CQI	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Bytes	0	3	3	6	11	15	20	25	36	39	50	63	72	80	93	93

A. Discontinuous Reception (DRX)

Under DRX[2], the UE periodically wakes up to monitor the PDCCH for a period of time, set by the *On Duration Timer* (ODT), in a cycle whose length and offset are called *Long DRX Cycle* (LDC) and *DRX offset* (DO) respectively. If scheduled during its *on* phase, the UE stays awake until either the ODT expires, or another timer, called *Inactivity Timer* (IT), expires, whichever occurs last. The IT is re-scheduled on each reception, and its purpose is to delay the *sleep* phase so that a burst of packets at the end of an *on* phase can be received correctly. Note that the IT must be *at least* one TTI, and that it *prolongs* the duty cycle without altering the cycle, as shown in Figure 3. Uplink retransmissions have priority over DRX timings, hence the UE must stay *on* at a TTI when an uplink H-ARQ process is in retransmission, whatever its resulting DRX status would be at that time. Moreover, during a RAC procedure, the UE must stay *on* until either a configurable maximum window has expired, or until the RAC request is responded to, whichever comes first.

Some traffic scenarios are characterized by periods of (possibly intermittent) traffic exchange, followed by little or no activity (e.g. web browsing). To handle these cases, another type of DRX Cycle – called the *Short DRX Cycle* (SDC) – has been defined. During inactivity periods, the cycle duration is given by the LDC. When the UE is *on* and is scheduled for a new transmission, it switches to SDC, i.e. to shorter cycles, for a number of consecutive times, known as *Short Cycle Timer* (*SCT*). The SCT is reset each time the UE is scheduled, hence the UE returns to LDCs after receiving no packets for SCT×SDT TTIs. Finally, the LTE standard allows the eNB to turn off the UE at any time. This is done via a *DRX-Command* MAC control element (*DCE*), i.e. a MAC header sent within a standard PDU. The latter stops *both* the ODT and the IT, thus sending the UE to sleep until the next wake-up time. If short/long cycles are configured, the SCT is restarted and the SDC will be used for the next cycles.

All the above parameters are configured through the Radio Resource Control (RRC) protocol. RRC signaling takes tens of TTIs and occupies downlink resources, which makes it infeasible for short-term adjustments. In other words, DRX configuration is not meant to cope with instantaneous queue length variations, rather it should be employed at larger timescales (i.e., seconds or more), comparable with flow lifetimes.

[2] The acronym DTX, which stands for *Discontinuous Transmission*, is sometimes used in the literature to refer to DRX in the uplink. In fact, there is only *one* mechanism in the standard, which goes by the name of DRX *and* affects both directions at the same time.

III. SIMULATION MODELS AND METHODOLOGY

In this section we describe the simulator that we use for our analysis, the relevant modeling assumption (i.e., the network and traffic models, and the UE power model), and the factorial analysis methodology.

A. Description of the simulator

Our evaluation is carried out using SimuLTE [35]-[36], a system-level simulator, comprising more than 40k lines of object-oriented C++ code, which includes all the layers of the protocol stack, from the physical to the application layer. Protocol layers and functions are conform to the Release 8 standard. SimuLTE has been developed for the OMNeT++ simulation framework [37]-[39]. The latter is a modular framework, which includes a considerable amount of network simulation models, notably INET [46], which boasts an impressive protocol matrix, all the TCP/IP stack, mobility, wireless technologies, etc. Furthermore, OMNeT++ allows one to keep a model's *implementation*, *description* and *parameter values* separate, and includes state-of-the-art debugging facilities (e.g. inspection of modules, animation of the flow of messages, etc.) and workflow automation tools (e.g., a manager for multiple runs in parallel, rule-based output data analysis, automated graphs, etc.). SimuLTE simulates the data plane of the LTE/LTE-A radio access network. It allows simulation of LTE/LTE-A in Frequency Division Duplexing (FDD) mode, with heterogeneous eNBs (macro, micro, pico etc.), using omnidirectional and/or anisotropic antennas, possibly communicating via the X2 interface [40]. Realistic channel models, fully 3GPP-compliant MAC, and resource scheduling in both directions are supported. In the current release, the Radio Resource Control (RRC) is not modeled, hence control messages traverse ideal channels.

SimuLTE implements eNBs and UEs as compound modules, as shown in Figure 4. These can be connected with each other and with other nodes (e.g. routers, applications, etc.) in order to compose networks. The simulator allows multiple TCP/UDP-based applications per UE. Each TCP/UDP App represents one end of a connection, the other end of which may be located within another UE or anywhere else in the topology. The IP module connects the *Network Interface Card* (*NIC*) to applications in the UE, whereas in the eNB it connects the eNB itself to other IP peers (e.g., a server running an application), via a PPP (Point-To-Point Protocol) connection. The *NIC* module implements the LTE stack, which includes:

- A PDCP-RRC module, which performs encapsulation and decapsulation and Robust Header Compression (ROHC)
- An RLC module, that performs multiplexing and demultiplexing of MAC SDUs to/from the MAC layer, and im-

plements the three RLC modes, namely *Transparent Mode* (TM), *Unacknowledged Mode* (UM) and *Acknowledged Mode* (AM), as defined in [40].

- A MAC module, where most of the intelligence of each node resides. Its main tasks are buffering of packets from upper (RLC) and lower layers (PHY), encapsulation of MAC SDUs into MAC PDUs and vice-versa, channel-feedback management, H-ARQ, DRX control, adaptive modulation and coding (AMC)
- A PHY module, that implements channel feedback computation and reporting, data transmission and reception, air channel emulation and control messages handling. It stores the physical parameters of the node, such as the transmission power and antenna profile (i.e., omni-directional or anisotropic). This allows one to define *macro-micro-, pico-*eNBs, with different radiation profiles.
- eNB scheduling in both the downlink and the uplink direction. In the uplink, all the mechanisms described in Section II are included. RAC collision probability is computed as a combinational problem, wherein any RAC preamble can be selected with the same probability by each UE [41], resulting in the following formula:

$$P_{collision} = 1 - \left(\frac{RACpre - 1}{RACpre} \right)^{RACreq},$$

where *RACpre* is the number of available RAC preambles, and *RACreq* is the number of RAC request in the current TTI.

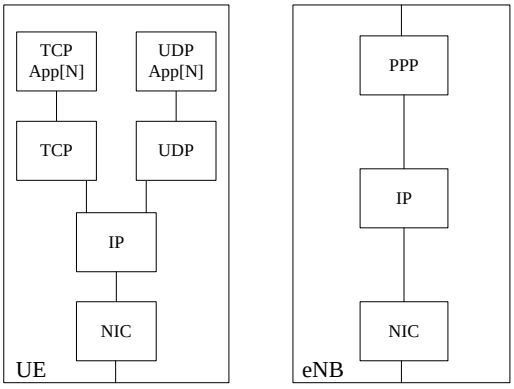

Figure 4 - UE and eNB module structure

B. Network model

The network used in the scenarios consists of a *core* network plus an LTE cell, as shown in Figure 5. For each communication, one of the endpoints is attached to the core network and the other is a UE in the cell. The core network adds a delay distributed as a Laplacian random variable (min 0 ms, mean 80 ms, max 120 ms), hence introducing jitter [43]. The LTE cell has an eNB equipped with an omnidirectional antenna at its center and UEs experiencing varying channel conditions. The main physical layer parameters are shown in Table 2. The RLC layer at the eNB is configured with the *Unacknowledged Mode*, with a fixed PDU size of 40 bytes. We use a realistic channel model with pathloss and fading: the former is based on the Urban Macro model (UMa) defined in [44], while for the latter we implement the Jakes model for Rayleigh fading [42]. UEs are dropped randomly within a square of a given size at the start of each run, then they move within it according to a Random Waypoint Model [45], at a speed uniformly distributed between 0 and 30 m/s.

We set the UE-to-eNB distance based on the channel and power model (see Table 3). More specifically, we use a "high" distance range (20 to 700m) for traffics which are downlink-only: this allows us to have a wider range for the CQIs. Conversely, we use a smaller distance range (10 to 500m) for uplink or bidirectional traffics. Given the UE power model, using the same range as for downlink transmissions would make correct reception at the eNB impossible for faraway UEs.

In order to analyze the system at sufficiently high loads while keeping the simulation overhead under control, we employ a spectrum of 10MHz with high-bandwidth applications (e.g., Video on Demand), and of 5MHz with low-bandwidth ones (e.g., VoIP). We expect full-spectrum simulations to yield qualitatively similar results, with due scale factors.

We employ two schedulers on the eNB side, namely MaxC/I and Proportional Fair (PF). The first one sorts backlogged UE by descending CQI (ties are broken by UE ID). This way, UEs with low CQI may be starved when the utilization is high, but the highest instantaneous cell throughput is always achieved. The second one sorts UEs by descending *PF score* $r_i / \overline{R_i}$, where r_i is the achievable rate at the current TTI (inferred by the UE's CQI), and $\overline{R_i}$ is the UE's historical rate, updated as $\overline{R_i} \leftarrow (1-\alpha) \cdot \overline{R_i} + \alpha \cdot r_i \cdot 1_{\{i \text{ is scheduled}\}}$. PF score combines channel conditions (given by the numerator) with waiting time priority (given by the exponential decay of the denominator), thus striking a balance between efficiency and fairness. We choose α equal to 0.05, following [50].

With both schedulers, UEs are served exhaustively in order of descending score, until no more UEs are backlogged or the frame is full. Both schedulers are made DRX-aware, meaning that they only schedule UEs in the *on* phase, but do not otherwise exploit energy efficiency considerations (e.g., by possibly prioritizing those UEs which are nearest to their sleep period). The scheduler type will be considered as a factor, so as to analyze possible interactions with DRX parameters. A comparative study of MAC schedulers specifically designed for DRX is left for future study.

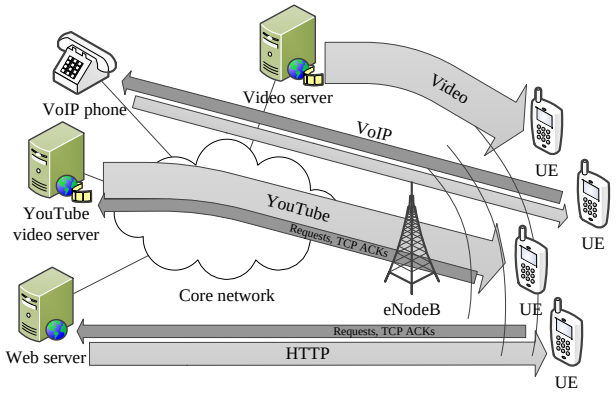

Figure 5 – Network model

TABLE 2 – PHYSICAL LAYER PARAMETERS

eNB Tx Power	40 dBm
eNB Noise Figure	2 dB
eNB Cable Loss	2 dB
UE Tx Power	24 dBm
UE Noise Figure	7 dB
Thermal Noise	-104.5 dBm

TABLE 3 – SCENARIOS

Traffic Type	**Distance Range**
Downlink only: VoIP DL, VoD	[20; 700] m
Uplink only or bidirectional: VoIP (UL, UL+DL), HTTP, YouTube	[10; 500] m

C. UE power model

As for the UE power model, we adopt the RF modem consumption model in [29], which further extends the one on which most of the related work mentioned in Section 0 is based [30]. It has three states and four transitions, each one with an associated power consumption, reported in Figure 6. The *LightSleep* state represents the RRC_CONNECTED state. It is used for short inactivity periods, when the UE powers down some of its circuitry. *DeepSleep* represents the RRC_IDLE state, used for longer inactivity periods wherein the UE powers down more hardware. In our simulations, applications are considered to be always active, hence the UE never enters the *DeepSleep* state. In the *Active* - *NoData* state the UE has the whole circuitry powered up but does not send/receive any data. In the other *Active* substates (i.e, RX, TX, RX+TX) the UE receives, sends, or receives and sends data from/to the eNB. Note that power consumption is different whether the UE is receiving, transmitting, or both. While the receiving consumption is fairly independent of the UE channel quality, the transmission one does depend on it, since a center-cell UE will use less power than a border-cell UE for the same PDU. The power consumption used in the model represents that of a border-cell UE.

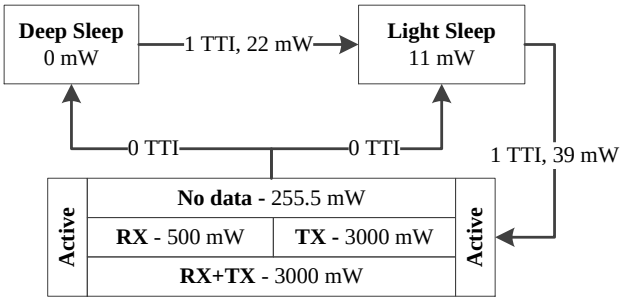

Figure 6 – Power consumption model

D. Application models

We describe in detail the models used for VoIP, Video on Demand, HTTP and YouTube applications.

1) Voice over IP

Voice over IP is modeled according to [32]. The employed codec is the GSM AMR Narrow Band (12.2 kbit/s) with VAD (no packets are sent during silences). The talkspurts and silence period durations are distributed according to Weibull functions, coherently with a one-to-one conversation model. Header compression is employed. The set of parameters is summarized in Table 4.

TABLE 4 – VOIP MODEL PARAMETERS

Talkspurt duration (Weibull distribution)	Shape scale	1.423 0.824
Silence duration (Weibull distribution)	Shape scale	0.899 1.089
Codec Type	GSM AMR Narrow Band (12.2 kbps) w. VAD	
VAD Model	One-to-one conversation	
Header Compression	Active (RTP+UDP+IP headers = 6 bytes)	
Packet length	32 bytes/frame + 6 bytes Hdr + 1 byte RLC	

As far as performance metrics are concerned, we compute the Mean Opinion Score (MOS) [34], which predicts the quality experienced by human users by combining losses and mouth-to-ear delays in a codec-specific formula. The MOS ranges from 1 (unintelligible) to 5 (perfect), and a MOS above a 2.5 threshold in at least 80% of the talkspurts is considered acceptable for the employed codec. Mouth-to-ear delays are accounted for by including the application layer, i.e. encoding/packetization delays and, more importantly, playout buffer delays and losses. Playout buffering is in fact a major source of delay and losses, and cannot be neglected. The receiver employs an *optimal* playout buffer [32], whose performance upper bounds that of any real-life playout buffer. The optimal buffer computes *a posteriori* the playout delay of each talkspurt that maximize the MOS for that talkspurt, hence being non-causal. As shown in [32], optimal buffering allows one to discount buffering-induced MOS degradations, while maintaining a good degree of realism at the same time. When analyzing bidirectional conversations, the activity in both directions are linked using the model in [33].

2) Video on Demand

Video on Demand (*VoD*) traffic is modeled by a streaming source that generates packets according to a pre-encoded MPEG4 trace file ([2]) whose parameters are summarized in

Table 5. The key performance metrics are *frame delay* and *frame loss*. Frame delay affects the amount of buffering required at the destination, as well as the initial playback delay. Since VoD is non-interactive, a shorter delay is preferable, but a higher delay may not heavily impair the user experience. On the other hand, frame loss *does* impair it, and heavily so, hence has to be kept very small. In MPEG4 video streams, frames are correlated, and some are necessary to decode others. For this reason, the frame type (I-frame, P-frame or B-frame), is carried in the packet, and losses are accounted for coherently (i.e., the loss of an I-frame determines the loss of the whole Group of Pictures (GoP) that relies on it for decoding).

TABLE 5 – VOD TRACE STATISTICS

Min frame size	26 Bytes
Max frame size	4686 Bytes
Mean frame size	266.759 Bytes
Mean bit rate	53.352 kbps
Peak bit rate	937.200 kbps
Frames per second	25

3) HTTP

The HTTP model simulates web traffic based on a set of CDF (Cumulative Distribution Function) data derived from live `tcpdump` traces. The communication is composed of page requests of fixed size, each one followed by one main object plus zero or more embedded objects. The delay between two consecutive page request is called *reading time*. The time between two consecutives object downloads is called *server response time*. The number of objects per page and their respective size is modeled using a truncated Pareto distribution and a truncated log-normal distribution (non-integer values are rounded up). The set of parameters is summarized in Table 6. The key performance indicator is the *page delay*, i.e. the time needed to receive a full page, including all the embedded objects, starting from the time the request is issued.

TABLE 6 – HTTP TRAFFIC MODEL

Reading Time [s] (exponential distribution)	Avg.	25
Objects per Page [#] (truncated Pareto distribution)	Avg. shape	6.64 2
Bytes per Object [byte] (truncated log normal distribution)	Avg. Std.	6.17 2.36
Request Size [byte]	constant	320
Response Time [s] (double exp. distribution)	Avg.	0.13

E. YouTube

YouTube traffic is modeled according to [24]. Each application instance is composed by a *video server* that streams data via a TCP connection to a *video client*. For each video, the server first sends an *initial burst*, corresponding to t_B seconds of video data, thus filling up the client buffer and buying some slack for possible future congestions. After the initial burst, a *throttling phase* starts, where data is sent in relatively large bursts (64 kB each) at a rate equal to $k \geq 1$ times the video playout rate. The client starts the playout after it collects γ packets [25]. When the buffer becomes empty, the client pauses and resumes when γ packets have arrived.

Work [26] measures the user QoE of a YouTube session, and relates it to the number N and length L of the playout pauses. The MOS formula for YouTube traffic is shown to be:

$$MOS = 3.5 \cdot e^{-(0.15L+0.19)\cdot N} + 1.5$$

Each UE has a dedicated YouTube server, to avoid muddying the waters with server congestion issues. A session is composed of MPEG4 videos being sent sequentially, spaced by a relatively small *inter-video* time. The video trace used is the same of the VoD example. A summary of the parameters is given in Table 7

TABLE 7 – YOUTUBE TRAFFIC PARAMETERS

Video Duration	Uniform [50,80] s
Inter-video interval	Uniform [1,2] s
t_B	40 s
k	1.25
γ	100 packets

F. Factorial analysis

As described in section II.A, the number of tunable DRX parameters, hence of simulation *factors*, is large. Moreover, their effect can be different depending on the metrics being analyzed. One possible approach is a *full factorial* analysis, i.e. performing a simulation for each possible combination of the values of the factors. With k factors, each one with N_i values, the number of simulation runs that are required is:

$$s = \left(\prod_{i=1}^{k} N_i\right) \cdot r \,,$$

where r is the number of replicas of a scenario, usually set based on the desired statistical accuracy. Number s clearly becomes forbiddingly large even with relatively few factors. Besides simulation *time*, which can be always be abated by employing more or more performing hardware, the amount of data that need to be analyzed quickly becomes unmanageable.

One way to reduce the value of s is $2^k \cdot r$ *factorial analysis.* For each factor, only the *extreme values* of the interval (i.e., the lowest and the highest) are considered. Thus, only the cross-product of the extremes has to be considered, which yields $s' = 2^k \cdot r$ □ s. Given one metric, under assumptions which can be tested a posteriori, factorial analysis produces a base value, representing the mean averaged through the whole set of measurements, and its 95% confidence interval. Moreover a pair of values for each factor and combination thereof, describing its *absolute* and *relative* impact on the given metric is reported. The former yields the absolute variation of the metric value due to the transition of a factor from the lower to the upper extreme. Specifically, a positive absolute impact implies that the metric increases between the extremes, and a negative value implies the opposite (though neither guarantee that the metric is *monotonic* with respect to that factor). The relative impact is a percentage describing how much a factor impacts on the variation of a metric compared to the others.

We show the method through a simple two-factor example, which however can be easily generalized, and we refer the interested reader to books on experiment design and perfor-

mance evaluation (e.g., [27]) for a more thorough exposition. Consider a metric of interest y, which depends on two factors *A* and *B*. If we define two variables,

$$x_n = \begin{cases} -1 & \text{factor } j \text{ is } low \\ +1 & \text{factor } j \text{ is } high \end{cases}, n = A, B,$$

we can regress on x_A, x_B with a non-linear model as follows:

$$y = q_0 + q_A \cdot x_A + q_B \cdot x_B + q_{AB} \cdot x_A \cdot x_B + e,$$

where q_0 represents the baseline value (i.e., the part of *y* that remains constant when factors are varied), q_A, q_B represent the absolute contribution of each factor, q_{AB} is the joint contribution, and e is the experimental error. Recall that we are replicating each of the 2^2 scenarios *r* times: this means that the result of each replica *j* is a $\mathbf{R}^{(2^2)}$-vector $\mathbf{Y}_j = \left[y_{1,j} ... y_{2^2,j} \right]$, and we can also define a vector of sample means $\mathbf{M} = m_1, ..., m_{2^2}$. The absolute contributions can then be computed as:

$$q_j = \left(\mathbf{S}_j \cdot \mathbf{M} \right) / 2^2,$$

where $\mathbf{S}_j$ is the *j*-th column of the following $2^2 \times 2^2$ *sign matrix*, (*j* subscripts are reported above each column):

$$\mathbf{S} = \begin{matrix} 0 & A & B & AB \end{matrix} \begin{bmatrix} +1 & -1 & -1 & +1 \\ +1 & +1 & -1 & -1 \\ +1 & -1 & +1 & -1 \\ +1 & +1 & +1 & +1 \end{bmatrix}$$

Each row corresponds to one of the 2^2 scenarios, and all the possible combinations of low and high values for the two factors *A* and *B* appear in the rows. Each column is used to compute one of the 2^2 absolute contributions: the first column indicates absolute contribution q_0, which is in fact the average of the metric among all the experiments and replicas. The second and third columns are used to compute the absolute contributions of factors *A* and *B*, whereas the last one, whose signs are computed by taking the products of the elements in *A* and *B* columns, is the one related to q_{AB}.

In order to compute *relative* contributions, we need to apportion the *total variation* to each factor or combination thereof, or to the experimental error (which counts as *unexplained variation*). The total variation is given by the Sum of Squares Total (SST), i.e. $SST = \sum_{i,j} (y_{i,j} - \mu)^2$, where μ is the mean value of the metric averaged across all the experiments and $y_{i,j}$ is the sample of the *j*-th replica of the *i*-th scenario. After a modicum of algebra, it can be shown that:

$$\sum_{i,j} (y_{i,j} - \mu)^2 = 2^2 \cdot r \cdot {q_A}^2 + 2^2 \cdot r \cdot {q_B}^2 + 2^2 \cdot r \cdot {q_{AB}}^2 + \sum_{i,j} e_{i,j}^2, \quad (1)$$

where $e_{i,j}$ is the *error* (or *residual*), i.e., the difference between the *predicted* and *observed* value of the metric at the *j*-th replica of the *i*-th scenario. The latter can be rewritten as $SST = SSA + SSB + SSAB + SSE$, where each Sum of Square addendum on the right-hand side matches the corresponding one in (1), and accounts for the variation due to factors *A*, *B*, *AB* jointly, and to errors. Ratios SSx/SST are in fact the relative contributions. The ratio SSE/SST is the *unexplained variation*. The method works under two assumptions, namely that the errors are statistically independent, and that they are normally distributed. These two assumptions can be verified *a posteriori*, using visual techniques. The independence assumption is usually verified by plotting a scatterplot of the residuals against the predicted responses. The plot should show no visible trend (e.g., ascending or descending) in order for the assumption to hold. However, if the residuals are at least one order of magnitude smaller than the predicted responses, then trends can be ignored altogether. The second assumption (normally distributed errors) can be verified by plotting the residual quantiles against those of a standard normal distribution in a so-called *Q-Q plot*. If the result is *approximately* linear, then the normality assumption holds. Significant deviations from a linear behavior may hint at the fact that the regression model is inappropriate for the task at hand, e.g. because the limit values are too far apart. Furthermore, it is desirable that unexplained variation be reasonably low, e.g., up to few percentage points.

We use factorial analysis to show which factors should be tuned in order to achieve the desired effect on a metric, using the tool described in [28] plus some trace-parsing code. Besides DRX parameters, we will include the scheduler type in the analysis as a binary factor (MaxC/I or PF), in order to assess possible interactions with DRX settings. For each analysis we will report a table describing the absolute and relative impact of the parameters on the system metrics, together with the unexplained variation. Every scenario has been verified *a posteriori* for correctness.

IV. Performance Analysis

We present here performance results related to the applications described in the previous section, along with guidelines on how to set the DRX parameters for each one. Except where specified otherwise, each simulation run lasts for 200s, with a warm-up time of 20s where statistics are not collected, and is replicated five times with different seeds. Applications are started at a random time uniformly distributed in [0,5]s.

As we will see, the system performance is affected by many factors, both *quantitative* (e.g. number of UEs, ODT, LDT, etc.) and *qualitative* (whether to activate the DCE message or not, whether to use semi-persistent vs. RAC-based uplink scheduling, etc.). Qualitative factors will often be analyzed separately, and – when appropriate – we will resort to *factorial analysis* to evaluate the impact of quantitative ones on *cell-averaged* metrics. We will also draw *scatterplots* to evaluate how per-UE metrics are spread around the average per-cell value.

A. VoIP

VoIP is inherently bidirectional. Now, the DRX affects both directions simultaneously, since it is regulates the activity of a UE. However, the factors that play a role in VoIP performance are different in the two directions, since the scheduling processes are independent and inherently *different*. For this reason, we will first analyze the downlink and the uplink separately and then show how the above analyses converge in the case of bidirectional connections.

1) Downlink

We analyze the *downlink* (DL) part of a VoIP communication (i.e. the flow having the UE as a sink). We first show a

feature which is common to all the traffic types, regarding the impact of the DO. The latter can – and should – be set so as to mitigate contention on each TTI as far as possible: a wise choice is thus to minimize the amount of UEs that compete for resources at any TTI, which can be obtained by minimizing the overlap of their *on* phases as follows:

$$DO_i = (DO_{i-1} + ODT) \bmod LDC$$

Such *Minimum Overlap* solution is compared with a *fixed* and a *random* DO schemes. The first one makes two groups, one with DO=0 and one with DO=LDC/2, whereas the second assigns the DO randomly when the UE joins the cell. Figure 7 is a scatterplot of the MOS of each UE (i.e., each UE corresponds to a dot), with 100 to 300 UEs, under the three above DO selection schemes, using MaxC/I scheduling (results with PF are similar). As the figure shows, the *fixed* solution leads to poor MOS performance, already with 100 UEs (hence is not considered at higher loads), while the *Random* and *Minimum Overlap* show better results. Note that while the *average* MOS value of the last two solutions is similar, UEs are slightly less scattered with *Minimum Overlap*, i.e., the performance is more predictable. This is common to all scenarios and traffics, hence we assume Minimum Overlap henceforth without explicitly repeating the analysis.

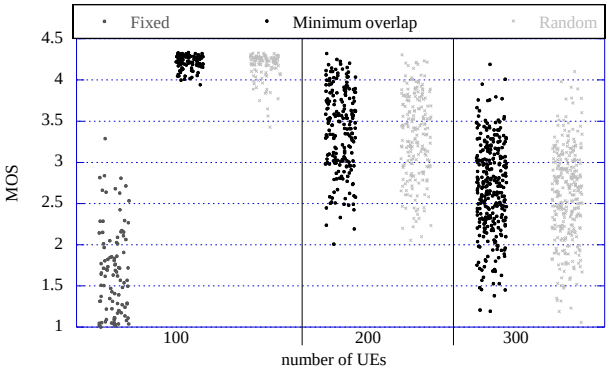

Figure 7 – MOS of VoIP conversation as a function of the no. of UEs for various DO selection strategies.

As far as quantitative factors are concerned, we can rule out *a priori* those related to alternating long/short cycles (i.e., SCT, SDC). In fact using long cycles to save power during downlink silence periods is of little impact, since *uplink* traffic will be transmitted during those (mutual silences being comparatively less frequent and much shorter than talkspurts). We are then left with analyzing the impact of ODT, LDC and IT together with the scheduling policy, which we do using factorial analysis with the parameter range in Table 8. Unlike with other traffics, the smooth nature of VoIP makes selecting the parameter range a straightforward application of common sense. Table 9 shows the impact of the three above parameters on the MOS, at both low (100 UEs) and high (300 UEs) loads. We first observe that the MOS decreases with the load, as resource contention does add a delay. The LDC has the highest impact on the MOS, and the impact is expectably negative. In fact, when the LDC is larger than the VoIP period, DRX further delays IP packets, as more than one IP packet is sent in an *on* phase, and this effect dominates the performance. As the load increases, however, the (positive) impact of the ODT increases. This is because at high loads the number of UEs competing for resources in a TTI is high, hence increasing the ODT increases the number of TTIs where a UE can be scheduled, thus improving its performance. The IT has a negligible impact at both low and high loads, as the VoIP is CBR during talkspurts, hence it is unlikely that prolonging the *on* phase on receipt of a packet will be of any use. Finally, the impact of scheduling on the MOS is minor. This recurring phenomenon, which apparently defies common sense, deserves an ad hoc explanation. Unless the network is in saturation, it makes perfect sense that the performance is largely dominated by DRX settings, and depends less on scheduling: in fact, the two schedulers sort backlogged UEs differently, but this makes no difference as long as every one of them will be scheduled in the current TTI, or soon enough as to make no matter. This of course does not imply that the scheduler has *never any effect* on system performance, especially in terms of perceived QoS. We will come back to the relationship between scheduling and QoS at the end of this section. Meanwhile, we observe that the above phenomenon occurs with all types of traffic, and the same explanation applies, hence we will omit restating it.

The impact of the four factors on the power is shown in Table 10, and exhibits a similar trend, with the ODT understandably having a higher impact (the power consumption is in fact proportional to the duty cycle ODT/LDC). The base value decreases with the load. As resource contention increases, in fact, UEs are scheduled less often, hence tend to receive more data in a single burst, which is more efficient from a power consumption point of view. Again, scheduling does not affect power consumption noticeably. The criteria to validate factorial analysis (i.e., Q-Q plots and error distributions) are met in this case, as well as for the other traffics where this technique is used, hence we will omit repeating this hereafter. Moreover, note that the unexplained variation is always small to negligible.

Given that traffic is CBR during talkspurts and consists of *short* packets, under reasonable LDC values it is hardly likely that more than one MAC PDU (itself possibly carrying more than one VoIP frame) will be received on each DRX period, barring severe jitter conditions. We can thus safely send an UE to sleep using DCE every time it is scheduled. This cuts down the *on* phase, whatever the ODT and IT values. DCE messages are piggybacked within a MAC PDU, hence have negligible to null cost in terms of occupied resources (most of the times they fit into bits that would otherwise be filled with padding). Figure 8[3] shows the power saved by using the DCE, in various configurations. Noticeable reductions are obtained even for ODT=1, since the IT is bypassed (recall that the IT cannot be null). The saving depends on the ODT, rather predictably, and decreases with the load. The latter effect is justified by the fact that a higher load implies a reduced chance of being scheduled

[3] Figures are drawn using MaxC/I as a scheduler, unless specified otherwise. Those with PF are always very similar, hence we omit showing them.

(and, thus, sent to sleep) *early* in the *on* phase. In Figure 9 we show the effects of the DCE on MOS for two load scenarios (100 and 300 UEs), two ODT (1ms, 10ms), two LDC (20ms, 80ms) with/without the DCE. The figure shows that the MOS is hardly affected at all by the DCE, some difference being observable for LDC=80. In this case, in fact, the DRX cycle is four times the period, making it highly likely that *more than one* VoIP packet will be available at the beginning of each *on phase*. If those packets are not transmitted all in the same TTI (possibly due to high contention, hence fewer available resources), the DCE may delay the remaining one(s)by one cycle, by sending the UE to sleep after the first one. However, even in that case, the MOS reduction is minor, *because the added jitter is easily absorbed by the receiver playout buffer*.

Summarizing the above, the practical guidelines for configuring DRX in downlink VoIP flows are the following:

- Scatter UE *on* phases using the DO, so that roughly the same number is active on each TTI;
- always use the DCE, and send UEs to sleep as soon as they are scheduled;
- set the LDC according to the desired target MOS, regardless of the cell load: a higher MOS is achieved using a smaller multiple of the VoIP frame period;
- increase the ODT with the *cell load* to compensate for a reduced scheduling probability.

TABLE 8 – PARAMETER RANGE FOR FACTORIAL ANALYSIS, DL VOIP

Name	Min Value	Max Value
ODT	1	10
LDC	20	80
IT	1	10
Scheduler	MaxC/I	PF

TABLE 9 – FACTORIAL ANALYSIS, DL VOIP MOS

	100 UEs		300 UEs	
Base Value	4.070		3.389	
95% Conf. Int.	± 0.0049		± 0.0076	
	Relative	**Absolute**	**Relative**	**Absolute**
LDC	70.02%	-0.256	73.82%	-0.465
ODT	15.34%	0.119	23.90%	0.264
ODT×LDC	8.85%	0.091	0.39%	-0.034
IT	2.34%	0.047	0.31%	0.030
Scheduler	0.01%	0.003	0.08%	-0.015
Other 10[4]	2.25%	-	0.71%	-
Unexplained	1.22%		0.94%	

TABLE 10 – FACTORIAL ANALYSIS, DL VOIP POWER CONSUMPTION [mW]

	100 UEs		300 UEs	
Base Value	7.71E+04		6.73E+04	
95% Conf. Int.	± 5.04E+02		6.55E+02	
	Relative	**Absolute**	**Relative**	**Absolute**
LDC	43.53%	-3.37E+04	42.71%	3.25E+04
ODT	33.24%	2.94E+04	37.70%	-3.05E+04
ODT×LDC	13.72%	-1.89E+04	15.62%	-1.96E+04
IT	7.36%	1.39E+04	2.64%	8.08E+03
Scheduler	0.08%	1.44E+03	0.00%	2.18E+02
Other 10	1.78%	-	0.52%	-

[4] In this and in some of the following tables, this line reports the total of the factors, or combinations thereof, whose *individual* contributions are negligible.

Unexplained	0.46%	0.82%

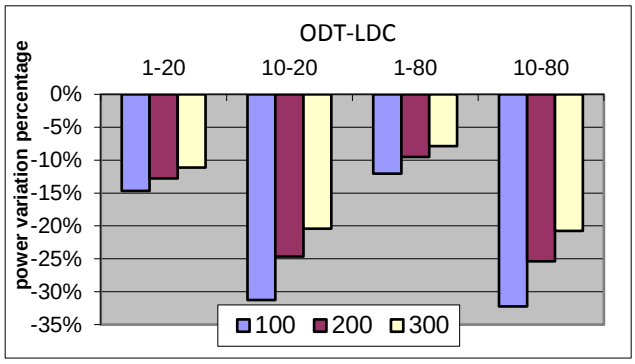

Figure 8 – Average power saving brought by DCE over a baseline DRX with the same parameters. ODT={1,10}, LDC={20,80}.

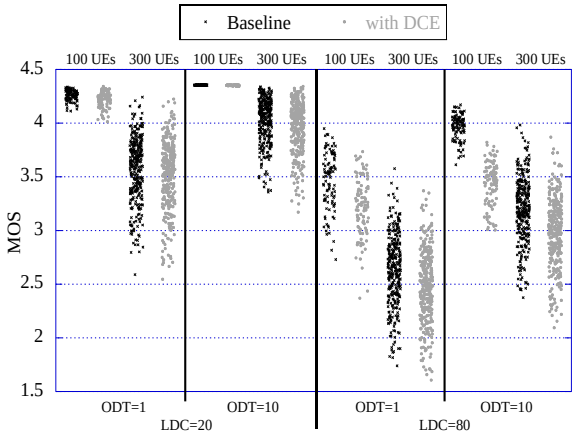

Figure 9 – MOS of downlink VoIP flows with baseline DRX (red) and DRX with DCE (blue). ODT={1,10}, LDC={20,80}. For each ODT-LDC pair the number of UEs is 100 (left) and 300 (right).

2) Uplink

In the uplink direction, packet generation *can* be assumed to be perfectly periodic, unlike for the downlink. The UE signals the arrival of new traffic to the eNB via RAC requests. RAC requests modify the DRX state as explained in Section II. For this reason, there is no point in using the DCE, and the ODT and the IT can be set to the minimum without any impact on the performance, which only depends on the LDC *and* the success probability of RAC requests. This makes factorial analysis redundant. The RAC success probability decreases with resource contention, i.e., with the cell load. Hence the uplink capacity depends heavily on the effectiveness of the RAC mechanism. Figure 10 shows the MOS of 50 to 250 UEs with several LDC values, using the three uplink scheduling strategies described in Section II, i.e. standard RAC, BW stealing and SPS. With standard RAC, the figure shows that increasing the LDC from 20ms to 80ms has a twofold effect. A higher LDC, in fact, delays packets, but it also decreases the rate of RAC requests, hence their contention, since more than one VoIP packet will be transmitted at the onset of each *on* phase, thus leaving more room for data transmission. The first (negative) effect is observable at low loads, whereas the second (positive) one prevails at high loads. Bandwidth stealing does increase the efficiency of the RAC mechanism: in fact, the MOS is generally higher, more so at higher loads, when saving the uplink resources otherwise occupied by BSR transmission becomes significant. SPS is instead *inefficient* at the cell capacity level, since it books resources for the long term, based on the channel conditions at the *onset* of a talkspurt. In fact, at the beginning of a talkspurt the UE issues a RAC request, and the eNB uses the CQI measured at that time to serve the subsequent requests. That CQI may of course be considerably worse than the average one for that UE in the rest of the talkspurt, whereas RAC-based scheduling always uses fresher CQIs. This inefficiency is multiplied by the number of VoIP packets that a periodic grant should accommodate, hence weighs more heavily with larger LDCs. While underestimating the CQI leads to wasting resources, overestimating it reduces the H-ARQ success probability, as shown in Figure 11, thus generating a larger number of retransmissions. Figure 12 reports a comparison of *average* MOS values, normalized to those obtained using standard RAC in the same conditions, confirming that BW stealing brings significant benefits at high loads, and SPS reduces the MOS in all configurations.

On the other hand, the three scheduling mechanisms have an impact on power consumption. Figure 13 shows the power saving of BW stealing and SPS, with respect to the average consumption achievable with standard RAC. BW stealing always reduces power consumption, especially at lower LDCs (20ms), where UEs are highly likely to complete the transmission of a VoIP packet within one RAC handshake. SPS, on one hand, allows more conservative DRX configuration than BW stealing. In fact, an ODT of 1 is enough to cope with periodic grants in the steady state (i.e., after the beginning of a talkspurt), whereas RAC-based scheduling (even with BW stealing) requires UEs to stay on for 3 TTIs at least just to cope with the delay of the RAC replies (see the timings of Figure 2). This justifies the more pronounced power saving obtained with lower LDC values. On the other hand, with higher LDC values, the size of a VoIP burst increases, bringing a twofold negative effect: first, *larger* periodic grants are harder to fit in a frame, hence some UEs will fall back on using RAC anyway due to the lack of space in the frame. Those who do not, instead, will often experience a higher rate of retransmissions, due to the mismatch between the unsolicited CQI and the current channel conditions, a mismatch which increases with the number of VoIP frames being packed in a single grant. The above effects concur to increase the power consumption, thus reducing the benefits of using SPS.

Summarizing, the guidelines for the uplink are:

- Use BW stealing, and allow SPS only at low loads and with an LDC equal to the period.
- Set the LDC according to the desired MOS.
- Set the ODT and IT to one

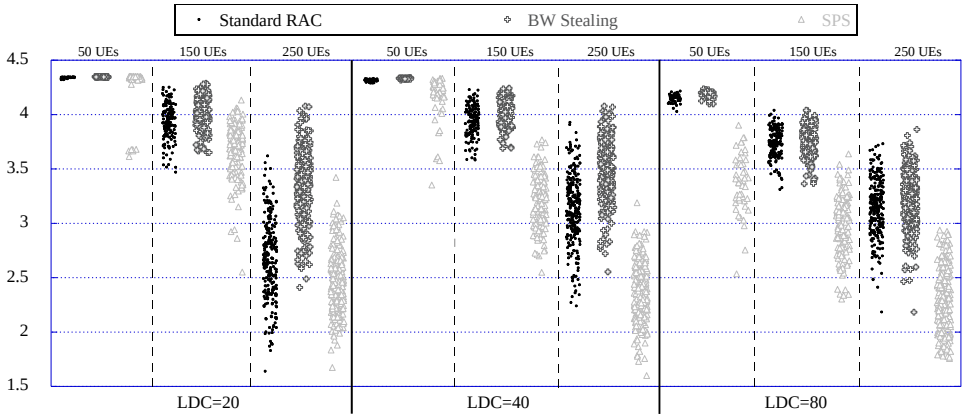

Figure 10 – UL MOS with standard RAC, BW Stealing and SPS. LDC={20,40,80}, UEs={50,150,250}.

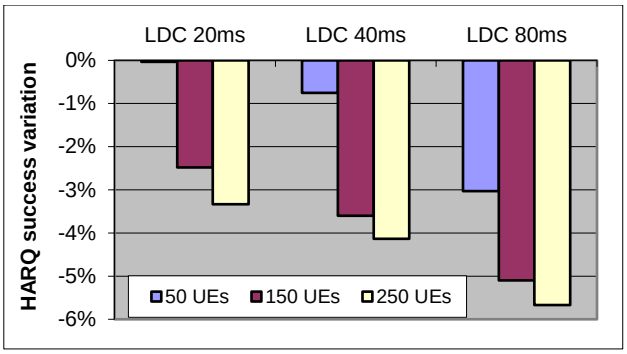

Figure 11 – Average HARQ success ratio reduction of SPS over BW Stealing.

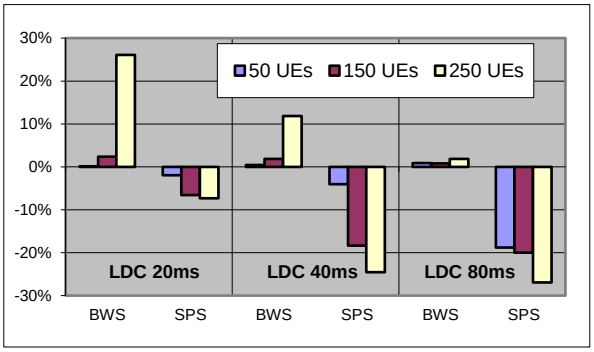

Figure 12 – Average UL MOS variation of BW Stealing and SPS compared to standard RAC-based scheduling.

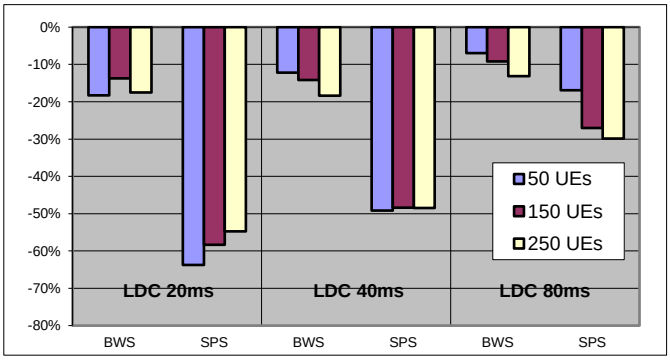

Figure 13 – Average power consumption variation of BW Stealing and SPS over standard RAC.

3) *Bidirectional VoIP*

In a VoIP communication, the same UE alternates between uplink transmission and downlink reception (mutual silences constituting a small enough fraction of the conversation time). DRX settings affect both directions, hence must be optimized for both. Luckily, the configuration guidelines in the previous two subsections are mutually compatible. Figure 14 – obtained with a cell radius of 500 meters, with DCE in the downlink and BW stealing in the uplink – shows that the *uplink* MOS decreases faster with the load: this is because uplink CQIs are generally lower than downlink ones for the same UE *in this scenario*. The above statement is in fact hardly general, since which direction acts as a bottleneck also depends on how cluttered the frame in that direction is. This in turn depends on the traffic mix, a safe bet being the downlink frame due to the asymmetry of the most popular applications. In Figure 15 we show the power consumed with bidirectional VoIP (green) and we compare it with the values obtained in downlink-only (red) and uplink-only (blue) communications. The consumption is dominated by the uplink, which – on one hand – requires more power during active transmission, and – on the other – generally requires more *on* time to complete because of the H-ARQ handshake.

In Table 11 we report an example of the power saving that can be achieved when activating DRX, compared with its cost in terms of MOS variation. We can achieve high savings with a negligible decrease in terms of MOS. Better yet, activating DRX together with BW stealing can even *improve* the uplink MOS. This is because DRX – by using a *minimum overlap policy* – de-synchronizes UE RAC requests, thus increasing their success rate (the increase grows with the load, from 7% for 150 UEs to14% with 250 UEs).

TABLE 11 - POWER AND MOS VARIATION WHEN ACTIVATING DRX (IT=1, ODT=1, LDC=20)

UEs	Power	MOS DL	MOS UL
50	-80.49%	-0.12%	-0.16%
150	-76.62%	-0.11%	4.59%
250	-75.66%	-0.13%	11.12%

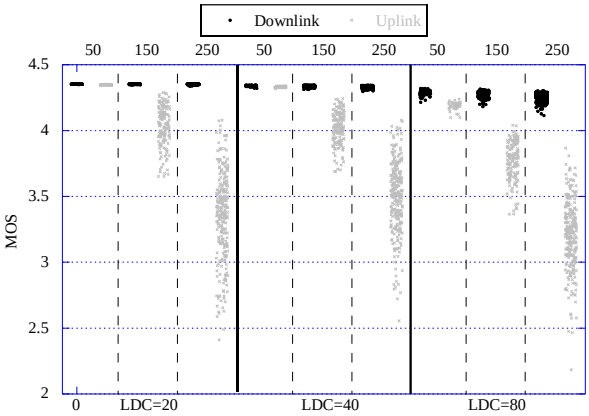

Figure 14 - MOS for Downlink and Uplink VoIP. ODT=1, IT=1.

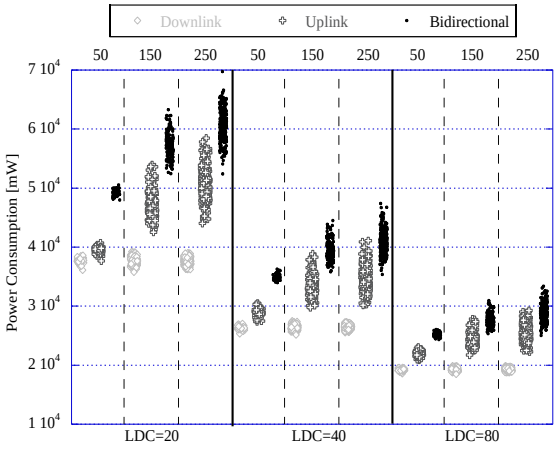

Figure 15 – Power consumption for DL, UL and bidir. VoIP. ODT=1, IT=1.

B. *VoD*

VoD has generally a higher bandwidth than VoIP. It has

constant inter-packet times at the source (subject to jitter induced by the core network, of course) and variable-size frames, leading to bursty traffic when large frames occur.

As for VoIP, we rule out *a priori* the DRX parameters related to alternating long/short cycles (SCD/SDT): in fact, VoD arrivals are periodic (barring jitter) and continuous, hence there is no point in using variable cycle durations. Moreover, the fact that arrivals are bursty implies that – unless onerous techniques such as deep packet inspection are used, and in conjunction with special coding techniques – it is generally harmful to send a UE to sleep on receipt of a PDU, since a large video frame is likely to occupy more than one PDU and to be transmitted over several consecutive TTIs. Hence DCE will not be considered. This leaves us again with four parameters (LDC, ODT, IT, scheduler) whose performance impact on frame loss, frame delay and power consumption we must evaluate. Regarding the quantitative ones, selecting the parameter range for factorial analysis bears some considerations. Since frame loss is the single most relevant QoE metric, and it exhibits a strong threshold behavior (i.e., poor quality beyond 1%-2% loss ratio), it makes sense to consider only the DRX configurations that meet that requirement. We experimentally observe that the frame loss is mainly due to the *duty cycle*, i.e. to the ratio ODT/LDC, rather than the value of the ODT and LDC parameters in isolation. At low duty cycles (i.e., smaller than ¼), the frame loss is unacceptably high. Rather counterintuitively, the reason is the poor level of *multi-user diversity*. In fact, the number of video flows that can be supported in our scenario is relatively small (the performance drops around 40 UEs, corresponding to an average frame utilization of 65%). With low duty cycles, it is highly likely that *only one UE* is *on* in a TTI. Therefore, when a large frame is transmitted, especially to a UE with a low CQI, a TBS as large as the whole frame will be allocated to that UE, something which is confirmed by the distribution of the TBSs recorded during the simulations. This happens with both schedulers, as they only differ in the way they sort backlogged UEs. Now, the maximum number of four H-ARQ retransmissions (despite soft combining), proves to be insufficient when TBSs are *very large*, because CQIs are reported so as to obtain a *block* error rate (BER) of 10%, hence express a per-Resource Block (rather than per-Transmission Block) error probability [5]. This effect can be quantified in a seemingly small 2-5% decrease of MAC transmission success probability, which is however amplified at the application layer by the fact that VoD frames are interdependent, hence the loss of a key frame affects a whole GoP. As a consequence, at low duty cycles we experience a high loss rate, despite having no buffer overflows at both the eNB and the UEs. Moreover, with low duty cycles, several video packets (which usually would arrive with some inter-packet delay due to network jitter) may build up a burst at the eNB simply by waiting for the next *on* phase, reinforcing the above effect. When the duty cycle increases, instead, two things happen simultaneously: large frames are fragmented into packets that can be transmitted in different TTIs of the same *on* window, on one hand, and MaxC/I favors UEs with higher CQIs, which tend to occupy a *part* of the frame, thus further splitting the transmission of potentially long TBSs over subsequent frames. This is confirmed by a recorded reduction of the TBSs, and – consequently – of the frame loss ratio.

We briefly speculate that one way to mitigate this phenomenon might be to adopt a different scheduling policy, notably one that limits the maximum number of RBs for a single UE in a TTI to a suitably small figure[5]. Such analysis (which is all but straightforward, given the obvious downsides of artificially limiting UE rates) is however outside the scope of this paper, and is left for further study.

Based on the above preliminary analysis, we perform our factorial analysis on the following three parameters: the *ODT Duty Cycle* (ODTD), together with the usual IT and LDC. This serves two purposes: on one hand, it allows us to capture the causes of variation more accurately, as we will show later on. On the other hand, it makes it easier to exclude from the analysis the region of the parameter space where performance (namely, the frame loss ratio) is unacceptably poor. Keeping that region in, in fact, would simply muddy the waters. We set the parameter range as in Table 12.

Table 13, Table 14 and Table 15 show the impact of DRX parameters on the frames loss, frame delay and power consumption, respectively. We first observe that the metrics are quite insensitive to the cell load, the only exception being the frame loss, and slightly so even then. As anticipated, the ODTD has by far the highest impact on the frame loss, with the LDC having only a minor effect. The impact of the LDC on frame delay is instead higher, especially for LDC=80ms as in that case a fixed delay of 40ms is introduced by DRX. Finally power consumption is affected only by the ODTD, which is expectable. In Figure 16 we analyze the trade-off between frame loss and power consumption. LDC is kept under 40ms as it has a negative impact on both loss and delay, with no benefits in terms of power consumption. The lower right part of the graph shows a low-power region (continuous-line cluster), characterized by ODTD equal to 2/4: in this case we should use an LDC equal to half the VoD period, as doing otherwise affects drastically the frame loss. In the left part we have a high-power region (dashed-line cluster), with frame losses under 1%. Note that in this case the metrics are roughly insensitive to the LDC.

In conclusion, the duty cycle should be set to at least 50% to guarantee reasonably low frame losses. This sets a firm lower bound to the power saving that can be achieved using DRX with VoD traffic. The IT reduces losses, but it is effective only at lower duty cycles. The LDC should not be increased beyond the video frame period.

TABLE 12 – PARAMETER RANGE FOR VOD FACTORIAL ANALYSIS

Name	Min Value	Max Value
ODTD	¼	¾
LDC	20	80

[5] Note that trying to reduce the TBS error probability by reducing the CQI alone (thus making the transmission more robust) would not be beneficial, since it would i) decrease the system capacity, and ii) increase the size of the TBS even more for the same amount of payload, thus defeating its very purpose.

IT	1	10
Scheduler	MaxC/I	PF

TABLE 13 – FACTORIAL ANALYSIS, VOD FRAME LOSS

	10 UEs		40 UEs	
Base Value	5.70%		7.52%	
95% Conf. Int.	± 0.17%		± 0.08%	
	Relative	**Absolute**	**Relative**	**Absolute**
ODTD	70.09%	-4.74%	70.83%	-6.16%
LDC	14.95%	2.19%	16.76%	2.99%
ODTD×LDC	8.02%	-1.61%	8.47%	-2.13%
IT	1.22%	-0.63%	1.60%	-0.92%
Scheduler	0.01%	0.07%	0.01%	-0.07%
Other 10	5.69%	-	1.75%	-
Unexplained	0.04%	-	0.60%	-

TABLE 14 – FACTORIAL ANALYSIS, VOD FRAME DELAY [S]

	10 UEs		40 UEs	
Base Value	9.31E-02		9.96E-02	
95% Conf. Int.	± 1.67E-04		± 2.34E-04	
	Relative	**Absolute**	**Relative**	**Absolute**
ODTD	51.41%	-6.20E-03	53.29%	-7.58E-03
LDC	28.21%	4.59E-03	26.51%	5.35E-03
ODTD×LDC	18.60%	-3.73E-03	16.76%	-4.25E-03
Scheduler	0.00%	-3.80E-05	0.07%	-2.71E-04
Other 11	1.76%	-	1.11%	-
Unexplained	0.02%	-	2.40%	-

TABLE 15 – FACTORIAL ANALYSIS, VOD POWER CONSUMPTION [MW]

	10 UEs		40 UEs	
Base Value	1.56E+05		1.56E+05	
95% Conf. Int.	± 8.03E+02		± 5.80E+02	
	Relative	**Absolute**	**Relative**	**Absolute**
ODTD	94.87%	5.71E+04	94.94%	5.70E+04
IT	1.71%	7.67E+03	1.73%	7.69E+03
Scheduler	0.02%	7.96E+02	0.00%	2.94E+02
Other 12	2.56%	-	2.87%	-
Unexplained	0.88%	-	0.46%	-

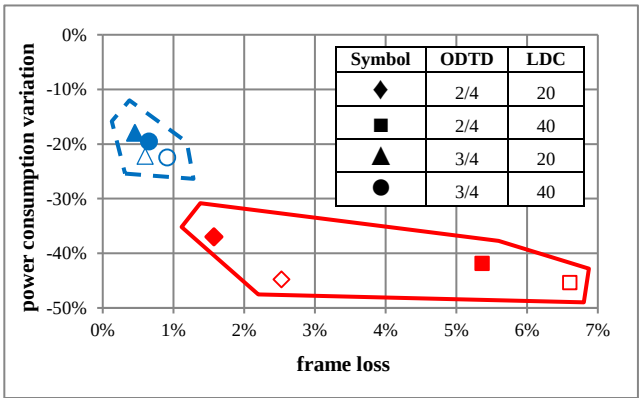

Figure 16 - Power saving vs. frame loss for VoD. Hollow markers are for IT=1, solid ones are for IT=10.

C. HTTP

HTTP is characterized by small packets in the UL (page requests) and bursts in the DL (object downloads). Long periods of inactivity (of the order of seconds) alternate with bursts of resource requests, again lasting seconds, involving both directions asymmetrically. In fact, the mechanism of short/long cycles has been envisaged to cope with these situations, hence this time we configure DRX with SDC and LDC. This implies that we must also set the number of short cycles (i.e., the value of SCT). It is intuitively clear that the SCT value per se is not very meaningful: what is meaningful, instead, is the product $SCT \times SDC$, which determines the time at which the "long cycle" regime resumes after a packet arrival. Therefore, similarly to what we have done for video, we define we define the Cumulative SCT (CSCT) as $CSCT = SCT \times SDC$ and use the latter in the factorial analysis. We avoid using DCE for the same reasons explained for VoD. Since the size of HTTP requests is not known in practice (although it is constant in the simulator), we refrain from using BW stealing in the uplink hereafter. This is however a minor detail, given the strongly asymmetric nature of this traffic.

The parameter range for factorial analysis is listed in Table 16. Given the sporadic nature of HTTP traffic and its intrinsic unpredictability, we empirically derived the values of the previous parameters trying to obtain non overlapping intervals for each parameter combination while exploring as many values as possible. Table 17 and Table 18 report the impact of all the DRX parameters on the power consumption and the page delay, respectively, for two load scenarios. We observe that the impact changes weakly with the load, and that – for this type of traffic – the delay performance depends on many factors simultaneously. Those that have the highest impact are the two cycle lengths, the SDC roughly double as much as the LDC. Moreover, the IT has the maximum impact (compared to VoIP and VoD), and the CSCT factor ranks fourth.

The two parameters that most impact power consumption are ODT and LDC (and, possibly, their ratio, since their combined impact is high as well): in fact, the duty cycle determines the power consumed during inactivity periods, which make most of the simulation time. During page download, instead, the regime is dominated by the SDC timer, hence the ODT/SDC ratio represents the duty cycle during activity periods, however, this has a negligible impact on the power consumption. Thus, the ODT/SDC ratio should be kept high to have low page delays. The IT can be kept high, as it has a very low impact on power consumption, whereas it decreases the page delay. Finally, note that the "unexplained" percentage is non-negligible. This is because HTTP traffic is less predictable than the other two, hence a larger spread among the various UE (given by, e.g., the different channel conditions) is bound to show, especially at lower loads.

TABLE 16 – PARAMETER RANGE FOR HTTP FACTORIAL ANALYSIS

Name	Min Value	Max Value
ODT	1	10
LDC	160	2048
IT	1	10
SDC	20	80
CSCT	40	320
Scheduler	MaxC/I	PF

TABLE 17 – FACTORIAL ANALYSIS, HTTP PAGE DELAY [S]

	50 UEs	100 UEs
Base Value	7.167	7.058

95% Conf. Int.	± 0.071		± 0.051	
	Relative	**Absolute**	**Relative**	**Absolute**
SDC	24.88%	1.779	24.35%	1.691
LDC	13.32%	1.302	13.81%	1.273
IT	9.07%	-1.074	10.13%	-1.091
CSCT	6.50%	-0.909	7.64%	-0.947
ODT×SDC	6.11%	0.882	6.42%	0.868
ODT	5.18%	-0.812	5.33%	-0.791
IT×SDC	4.33%	-0.742	5.04%	-0.769
IT×CSCT	4.28%	-0.738	4.35%	-0.715
LDC×CSCT	1.58%	-0.449	1.56%	-0.427
ODT×CSCT	1.49%	0.436	1.36%	0.400
Scheduler	0.01%	0.040	0.02%	0.048
Other 52	15.73%	-	15.84%	-
Unexplained	7.54%	-	4.20%	-

TABLE 18 – FACTORIAL ANALYSIS, HTTP POWER CONSUMPTION [mW]

	50 UEs		**100 UEs**	
Base Value	1.88E+04		1.89E+04	
95% Conf. Int.	± 1.89E+02		± 1.15E+02	
	Relative	**Absolute**	**Relative**	**Absolute**
ODT	39.25%	4.72E+03	40.30%	4.42E+03
LDC	25.23%	-3.79E+03	29.74%	-3.79E+03
ODT×LDC	16.91%	-3.10E+03	19.58%	-3.08E+03
SDC	1.83%	-1.02E+03	1.39%	-8.20E+02
ODT×SDC	1.32%	-8.65E+02	0.63%	-5.52E+02
IT	0.87%	7.04E+02	0.77%	6.11E+02
Scheduler	0.00%	-1.09E+01	0.69%	-5.76E+02
Other 56	2.83%	-	3.13%	-
Unexplained	11.77%	-	5.15%	-

Figure 17 shows the average power saving vs. the average page delay increase with respect to always-on UEs, in a number of DRX configurations with 100 UEs. Of the roughly 100 analyzed configurations, only those yielding a page delay increase smaller than 300% are shown. We can observe two clusters, characterized by two values of CSCT (320ms and 40ms). In general, a higher CSCT warrants a better trade-off between power and QoS (left cluster), all else being equal. In fact, the larger the CSCT, the more likely it is that subsequent objects of the same page are requested and downloaded within the same burst of short cycles, without having to pay the overhead for the next *long* cycle. Increasing the CSCT beyond 320ms yields diminishing returns, however: the delay does not decrease significantly, and the power consumption increases, albeit slowly. We thus omit to draw high-CSCT clusters for the sake of readability. A high LDC warrants low power consumption (during inactivity periods), but adds a delay at the onset of a new page request. Finally, varying the IT can be used for a finer tuning: its power cost is in general very low compared to its benefits in terms of page delay. Note that reducing the ODT further warrants very high delays (an increase larger than 300% in all configurations we analyzed).

Summing up, a power saving around 90% is achievable at the cost of increasing the delay less than 40%. The optimal configurations are with an ODT equal to 10, the SDC around 20ms, a rather high SCT (e.g., 8), and an LDC below 320ms.

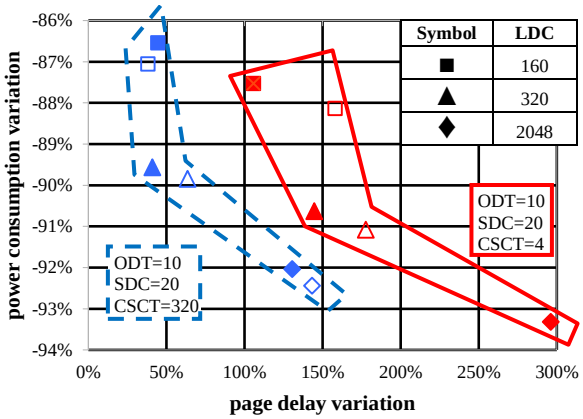

Figure 17 - Power saving vs. page delay increase. Hollow markers are for IT=1, solid ones are for IT=10.

D. YouTube

YouTube traffic is bidirectional and strongly asymmetric like HTTP. It has large bursts in the downlink (video transmission) with small packets in uplink (video requests and TCP ACKs). At the same time it has QoS constraints related to playout delays, similarly to VoD. For the same reasons as for HTTP we use RAC-based uplink scheduling. Even so, the uplink is never a bottleneck for this scenario. We simulate 400s of transmissions, to allow several videos per session, with desynchronized initial bursts.

Our analysis proceeds hierarchically: we first assess the importance of the duty cycle, which in fact determines the QoS, and then move to considering the further power saving opportunities warranted by long/short cycles. Since – as with VoD – we expect YouTube to request a high bitrate during video download, we set the limit values for factorial analysis in a fairly conservative way, as shown in Table 19. As we can see in Table 20, the main impact on MOS is given by the DRX duty cycle. However, the same duty cycle can be achieved differently at low and high loads: with 50 UEs (low load), the contention is low, hence a small ODT can be compensated for with a high IT. As the load increases, small ODTs become a hindrance, since a UE may not be scheduled during a short *on* phase, hence its IT is not triggered at all. This is consistent with what we observed, e.g., with VoIP. From an the energy point of view, setting the IT to higher values is a good choice, since it has a negligible impact on power consumption (Table 21). Note that, although we use the same trace file as for VoD, we are able to maintain an acceptable QoE for a higher number of users. This is because, on one hand, the size of the cell is smaller, as explained in Table 3, so as to guarantee high enough CQIs to the *uplink* leg of the traffic, and this yields higher *downlink* CQIs as well and increases the cell capacity. On the other hand, TCP does retransmit lost frames, hence reduces the frame loss ratio with respect to UDP-based VoD (especially, it avoids discarding correctly received frames due to missed dependencies). Finally, YouTube QoE degradations are due to playout pauses, and the MOS model elaborated in [26] degrades smoothly if pauses are short.

Figure 18 shows the trade-off between power and MOS var-

iation for 100 UEs, with respect to a baseline where DRX is not used. We initially keep ODT=10, and vary the IT and the LDC, obtaining two clusters. The first one (top-left cluster) represents high duty cycles, and is characterized by savings around 45% with negligible QoS degradation. Increasing the IT has almost no effect in this case. The second one (bottom-right cluster) has a low duty cycle, which yields sensibly higher savings, at the cost of degrading the QoS sensibly. Increasing the IT in this case somewhat improves the performance, without affecting power consumption, as suggested by the factorial analysis.

Having ascertained that the duty cycle dominates the activity phase, we take advantage of the sending pattern of YouTube to increase the savings. Figure 19 shows a snippet of traffic as seen from the client application. Both at the end of the initial burst and in the throttling phase, pauses in the order of several seconds can be observed. Clearly, this calls for alternating between *short* and *long* cycles. While the former regulate the duty cycle during activity phases, the latter are meant to intercept longer pauses. We set ODT=10 and SDC=20 to obtain a high duty cycle, and explore varying the CSCT and the LDC: our analysis, summarized in the bottom-left cluster of Figure 18, shows that the CSCT should be kept large enough to avoid reverting to the *long* cycle regime by mistake when an activity phase is instead ongoing. On the other hand, the LDC can be kept fairly large without affecting the QoE. By using this mechanism, we can achieve savings between 80 and 90% with little to none QoE degradation.

TABLE 19 – PARAMETER RANGE FOR YOUTUBE FACTORIAL ANALYSIS

Name	Min Value	Max Value
ODT	1	10
LDC	20	80
IT	1	10
Scheduler	MaxC/I	PF

TABLE 20 – FACTORIAL ANALYSIS, YOUTUBE MOS

	50 UEs		100 UEs	
Base Value	4.404		3.476	
95% Conf. Int.	± 0.018		± 0.014	
	Relative	**Absolute**	**Relative**	**Absolute**
IT	29.48%	0.305	13.74%	0.270
LDC	24.27%	-0.276	40.70%	-0.465
ODT	9.64%	0.174	40.70%	-0.465
IT×LDC	7.78%	-0.156	2.90%	0.124
ODT×IT	0.43%	0.037	1.70%	-0.095
Scheduler	0.32%	0.032	1.63%	-0.093
Other 9	0.62%	-	0.92%	-
Unexplained	4.89%	-	1.82%	-

TABLE 21 – FACTORIAL ANALYSIS, YOUTUBE POWER CONSUMPTION [mW]

	50 UEs		100 UEs	
Base Value	6.18E+04		6.15E+04	
95% Conf. Int.	± 1.84E+02		± 1.74E+02	
	Relative	**Absolute**	**Relative**	**Absolute**
ODT	51.51%	3.27E+04	50.65%	3.25E+04
LDC	29.27%	-2.46E+04	29.79%	-2.49E+04
ODT×LDC	18.95%	-1.98E+04	19.10%	-2.00E+04
Scheduler	0.00%	-2.39E+02	0.06%	1.15E+03
Other 11	0.19%	-	0.04%	-
Unexplained	0.08%	-	0.07%	-

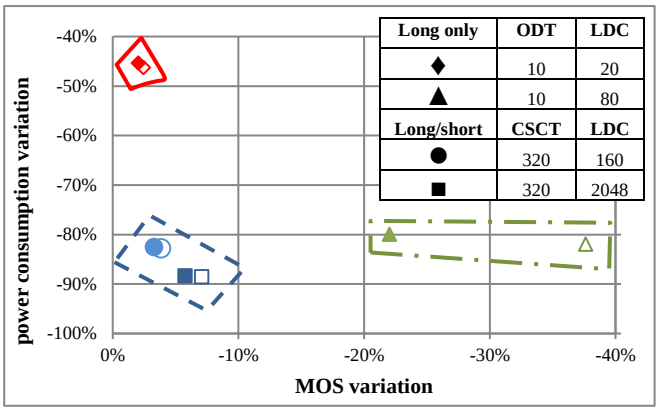

Figure 18 - Power saving vs. YouTube MOS variation. Hollow markers are for IT=1, solid ones are for IT=10. The top-left and bottom-right clusters are with *long cycles* only, the bottom-left cluster is with *long and short cycles* simultaneously, using ODT=10 and SDC=20.

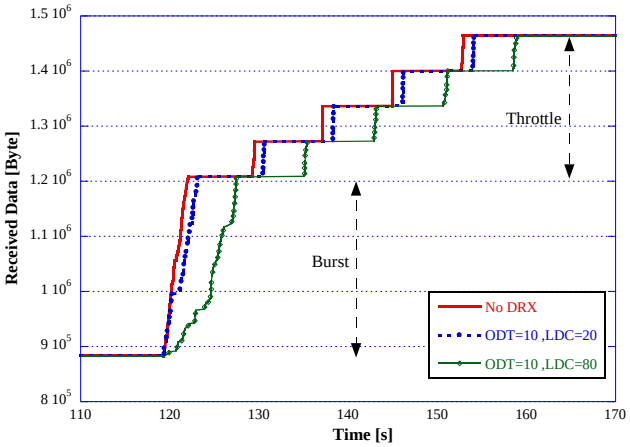

Figure 19 – Received YouTube data over time, IT=1, long cycles only.

E. On the impact of scheduling

In the previous sections we analyzed the effects of various DRX and system parameters on QoS and power consumption. As already discussed, the scheduling policy always had little impact on the performance. This does not imply that the scheduling algorithms do not affect QoS. Rather, the factorial analysis showed that the impact of scheduling is *minor with respect to* DRX parameters, within the limits of the analyzed scenarios. In other words, in all the cases we analyzed, selecting PF over MaxC/I (or vice versa) would not make up for DRX misconfigurations. On the other hand, once we tune the DRX parameters to a satisfactory trade-off between power

saving and QoS, we can appreciate the effects of scheduling on the QoS. As a first example, let us consider the downlink VoIP scenario presented in section IV.A. We set DRX parameters so as to obtain high performance in spite of contention, i.e., LDC=20 and ODT=10. Figure 20 shows the VoIP MOS obtained in this scenario for two different scheduling policies, namely MaxC/I and PF. The figure shows identical performances at low load (100 UEs). In this case, the DRX settings are indeed conservative for such a low load, thus any reasonable scheduling policy can be expected to yield the same performance. As the load increases (200 and 300 UEs) PF behaves more fairly: although the mean values are similar, PF exhibits a smaller variation than MaxC/I, as well as a slightly higher mean value. Finally, when saturation is approached (450 UEs), the opportunistic behavior of MaxC/I starts paying off, resulting in half a point of average MOS over PF. This is due to the fact that – by scheduling UEs with higher CQIs – MaxC/I occupies fewer RBs for the same load, hence saves more space for low-CQI UEs (recall that VoIP is CBR during talkspurts).

A similar trend can be observed with YouTube traffic, as shown in Figure 21. In that case, we use only *long* cycles, with relatively high duty cycles. With 125 UEs (i.e., a saturated network), MaxC/I leaves PF behind by half a point of MOS.

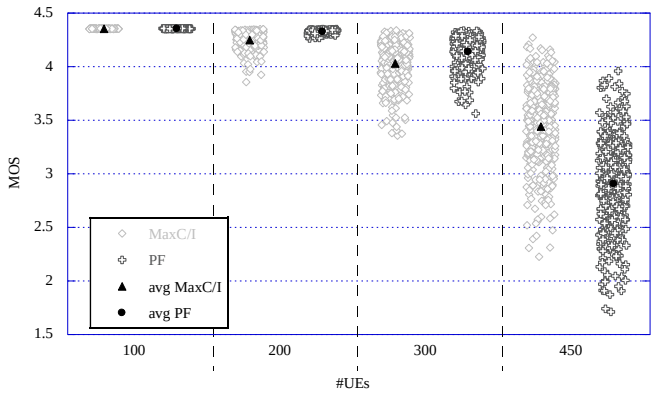

Figure 20 - Scheduling algorithm comparison with VoIP downlink traffic. ODT=10, LDC=20, IT=1. DCE is disabled.

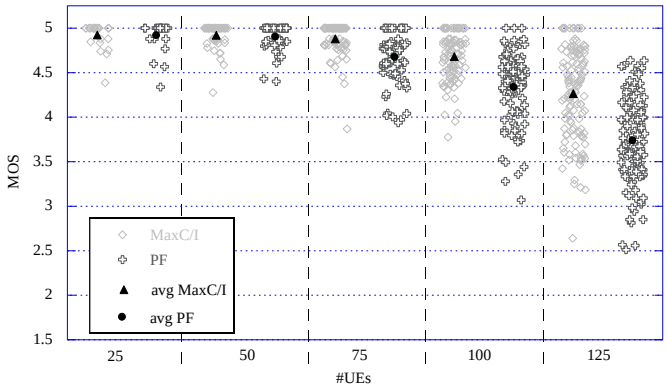

Figure 21 - Scheduling algorithm comparison with YouTube traffic. ODT=10, LDC=20, IT=10.

V. Related Work

A large number of papers on DRX for LTE have appeared recently, mostly in conferences and sometimes in journals. Some of them propose DRX-based solutions, i.e. scheduling ([6]) or extensions for newer LTE deployments, e.g., Carrier Aggregation [7] or TTI-bundling [8], hence are only marginally related to the object of this paper. Works on DRX *evaluation*, instead, such as [9]-[29], deal with one or more of the following:

1. Modeling DRX using analytical techniques (e.g., Markov or Semi-Markov) ([9]-[17]);
2. proposing adaptive techniques for setting some DRX parameters (e.g., [18]-[20]);
3. evaluating the performance of VoIP or HTTP traffic under DRX (e.g., [18]-[29], [11]).

Modeling accuracy and analytical tractability are contrasting requirements. Unfortunately, works that propose analytical models to evaluate the performance of DRX (e.g., [9]-[17]) do not compare their results to those that could be obtained when all the features of LTE are modeled (e.g., via simulation), which would allow a reader to appraise the extent of their accuracy. As anticipated in the Introduction, we believe that modeling applications running through LTE is a complex task, which defies analytical modeling.

Some works do exploit simulation to investigate the DRX performance (e.g., [21]-[23]). While simulation models (ours included) are always obtained under abstractions and simplifying assumptions, a detailed one can be expected to incorporate a higher number of features than most (tractable) analytical models. For instance, the non-negligible impact of the LTE protocol stack, complete of fragmentation, H-ARQ, physical channels and resource contention, is taken into account in the above works. However, these works only deal with VoIP in the downlink.

Regarding application models, most studies are carried out with Poisson traffic (e.g.,[10],[14],[15]), which lends itself to analytical tractability. Works studying HTTP traffic (e.g., [11], [19]-[20]) consider only its downlink leg. Works modeling VoIP (e.g., [21]-[23], [29]) normally place the VoIP sender directly at the eNodeB. By doing this, they neglect *jitter*, which is instead induced by the remote access network and the core network. Jitter in turn plays against DRX performance (as for both MOS and power consumption), since when an *on* phase is missed some power is wasted and a two-frame burst is likely to be created at the subsequent cycle.

Works that propose configuration of DRX parameters focus chiefly on long/short cycles: for instance, to the best of our knowledge, none consider *de-synchronization* of DRX cycles through *DO* selection, which plays a fundamental role in preserving cell capacity. Few investigate adapting the *on duration*, which is instead fundamental (e.g., to preserve VoIP QoS). None, finally, investigate using *DCE messages*, whose saving potential is indeed significant with VoIP. Finally, to the best of our knowledge, no systematic study based on factorial analysis has been attempted regarding DRX so far.

Recently, another type of applications, going by the collec-

tive name of *machine-to-machine* (M2M) traffic, have been gaining attention. The LTE network is an attractive choice to support this kind of applications, due to its ubiquitous access and built-in security. These applications are often characterized by very low bandwidths, i.e., tiny packets (normally fitting one RB) sent over periods of tens of seconds or more [47]. For these applications – which surely benefit from the power saving opportunities offered by DRX – configuring DRX parameters is hardly an issue at all: in fact, given the above traffic profile, any *reasonable* configuration will achieve huge power savings, at the price of little, if any, QoS degradation (i.e., a modest delay increase and a near-zero packet loss due to missed transmission opportunities) [48].

For M2M applications, instead, other problems are preeminent, such as maintaining synchronization over long sleeping periods where the UE enters a *DeepSleep* state [49], avoiding the excessive signaling due to many tiny connections by using *concentrators*, i.e., gateways that proxy a large numbers of such connections to the LTE network (possibly performing other functions, such as data aggregation or filtering). For these reasons, to avoid stating the obvious, we omitted discussing M2M applications within the framework of this paper.

VI. Conclusions and Future Work

In this paper we have analyzed the effect of DRX on the QoS and power consumption of UEs, with VoIP, VoD and HTTP traffics. The evaluation has been carried out by simulation, analytical modeling being out of the equation due to the intricacies of the LTE environment, using a statistically rigorous method called *factorial analysis*. For each type of traffic, an analysis of the impact of *qualitative* and *quantitative* factors related to DRX has been performed. This allowed us to identify guidelines for DRX configuration at the eNodeB in order to achieve the best QoS/power trade-off. In general, this trade-off appears to be favorable, meaning that high savings can be obtained with little to none QoS degradation, especially with TCP-based services.

As far as future work is concerned, the present one has given the schedulers for granted, whereas our preliminary results show that a DRX-oriented scheduler might further improve the balance between power consumption and QoS. We are actively pursuing this line of research at the time of writing.

Acknowledgements

The authors would like to thank Dr. Carlo Vallati of the University of Pisa for providing useful suggestions on factorial analysis.

References

[1] http://www.3gpp.org

[2] http://www-tkn.ee.tu-berlin.de/research/trace/pics/FrameTrace/mp4/index6e27.html

[3] H.Holma, A.Toskala, "LTE for UMTS - OFDMA and SC-FDMA Based Radio Acces", John Wiley & Sons, 2009

[4] Medium Access Control (MAC) Protocol Specification, 3GPP TS 36.321 (rel 9)

[5] 3GPP - TS 36.213, Physical Layer Procedures

[6] H. Bo, T. Hui, C. Lan, Z. Jianchi, "DRX-Aware Scheduling Method for Delay-Sensitive Traffic", IEEE Communications Letters, Vol. 14, No. 12, December 2010

[7] C.Zhong, T. Yang, L. Zhang, J. Wang, "A new discontinuous reception (DRX) scheme for LTE-Advanced Carrier Aggregation Systems with Multiple Services", in Proc. of VTC Fall 2011, Sept 5-8 2011, San Francisco, CA.

[8] S. Fowler, "Study on Power Saving Based on Radio Frame in LTE Wireless Communication System Using DRX", 2011

[9] L. Zhou, H. Xu, H. Tian, Y. Gao, L. Du, L. Chen, "Performance Analysis of Power Saving Mechanism with Adjustable DRX Cycles in 3GPP LTE", IEEE 68thVehicular Technology Conference, VTC Spring, September 2008, pp. 1-5

[10] V. Mancuso, S. Alouf, "Analysis of power saving with continuous connectivity", Computer Networks 56 (2012), pp. 2481–2493

[11] V. Mancuso, S. Alouf, N. C. Fofack, "Analysis of power saving and its impact on web traffic in cellular networks with continuous connectivity", Pervasive and Mobile Computing, 2012

[12] K. De Turck, S. De Vuyst, D. Fiems, S. Wittevrongel, H. Bruneel, "Performance analysis of sleep mode mechanisms in the presence of bidirectional traffic", Computer Networks, 56(10), July 2012, pp. 2494-2505

[13] Y. Wen, J. Liang, K. Niu, W. Xu, "Performance Analysis and Optimization of DRX Mechanism in LTE", in Proc. of IC-NIDC 2012, Bejing, China, 2012

[14] S. Jin, D. Qiao, "Numerical Analysis of the Power Saving in 3GPP LTE Advanced Wireless Networks", IEEE Trans. On Vehicular Technology, Vol. 61, No.4, May 2012

[15] K. Zhou, N. Nikaein, T. Spyropoulos, "LTE/LTE-A Discontinuous Reception Modeling for Machine Type Communications", IEEE Wireless Communication Letters, 2012

[16] S. Baek , B. D. Choi, "Analysis of Discontinuous Reception (DRX) with bothDownlink and Uplink Packet Arrivals in 3GPP LTE", in Proc. of QTNA 2011, Aug. 23-26, 2011, Seoul, Korea

[17] Y. Y. Mihov, K.M. Kassev, B.P. Tsankov, "Analysis and Performance Evaluation of DRX Mechanism for Power Saving in LTE", in Proc. of IEEE 26th Convention of Electrical and Electronics Engineers in Israel, 2010

[18] L. Liu, X. She, L. Chen, "Multi-User and Channel Dependent Scheduling Based Adaptive Power Saving for LTE and Beyond System", in Proc. of 16th APCC, 2010, Auckland, NZ.

[19] J. Wigard, T. Kolding, L. Dalsgaard, C. Coletti, "On the User Perfomance of LTE UE Power Savings Schemes with Discontinuous Reception in LTE", in Proc. of IEEE ICC'09, June 2009, Dresden, Germany

[20] T. Kolding, J. Wigard, L. Dalsgaard, "Balancing Power Saving and Single User Experience with Discontinuous Reception in LTE", in Proc. of ISWCS '08, October 2008, Reykjavik, Iceland

[21] K. Aho, *et al*., "Tradeoff Between Increased Talk-time and LTE Performance", in Proc. of ICN'10, April 2010, Menuires, France

[22] K. Aho, T. Henttonen, L. Dalsgaard, "Channel Quality Indicator Preamble for Discontinuous Reception", in Proc. of IEEE VTC Spring 2010, Taipei, TW

[23] K. Aho, T. Henttonen, J. Puttonen, L. Dalsgaard, T. Ristaniemi, "User Equipment Energy Efficiency versus LTE Network Performance", IARIA Int. J. on Advances in Telecommunications, vol 3 no. 3 & 4, 2010

[24] Ramos-munoz, J.J.; Prados-Garzon, J.; Ameigeiras, P.; Navarro-Ortiz, J.; Lopez-soler, J.M., "Characteristics of mobile YouTube traffic," Wireless Communications, IEEE , vol.21, no.1, pp.18-25, February 2014

[25] Staehle, B.; Hirth, M.; Pries, R.; Wamser, F.; Staehle, D., "YoMo: a YouTube application comfort monitoring tool", in Proc. of QoEMCS 2010, Tampere, Finland, June 9, 2010

[26] Hoßfeld, T.; Schatz, R.; Biersack, E.; Plissonneau, L.; "Internet video delivery in YouTube: from traffic measurements to quality of experience", In DataTraffic Monitoring and Analysis, Ernst Biersack, Christian Callegari, and Maja Matijasevic (Eds.). Springer-Verlag, Berlin, Heidelberg 264-301. 2013

[27] R. Jain, "The art of computer systems performance analysis", John Wiley & Sons, 1991

[28] C.Cicconetti, E.Mingozzi, C.Vallati, " A 2^kr factorial analysis tool for ns2measure", in Proc. of VALUETOOLS 2009, October 20-22, 2009, Pisa, Italy

[29] M. Polignano, D. Vinella, D. Laselva, J. Wigard, T.B. Sorensens, "Power savings and QoS Impact for VoIP Application with DRX/DTX Feature in LTE", in Proc. of VTC Spring, 2011

[30] Nokia, R2-071285,"DRX parameters in LTE", march 2007.

[31] 3GPP - TS 36.300 Evolved Universal Terrestrial Radio Access (E-UTRA) and Evolved Universal Terrestrial Radio Access Network (E-UTRAN); Overall description; Stage 2

[32] A. Bacioccola, C. Cicconetti, G. Stea, "User level performance evaluation of VoIP using ns-2", in Proc. of NSTOOLS'07, Nantes, France, Oct. 22, 2007.

[33] H.P. Stern, S. A. Mahmoud, K. Wong, "A comprehensive model for voice activity in conversational speech-development and application to performance analysis of new-generation wireless communication systems", Wireless Networks, Dec. 1, 1996

[34] ITU-T Recommendation G.107. The Emodel, a computational model for use in transmission planning. Dec. 1998

[35] SimuLTE website, https://github.com/inet-framework/simulte

[36] A. Virdis, G. Stea, G. Nardini, "SimuLTE: A Modular System-level Simulator for LTE/LTE-A Networks based on OMNeT++", in proc. of SimulTech 2014, Vienna, AT, August 28-30, 2014

[37] OMNeT++. http://www.omnetpp.org [accessed December 2013]

[38] Varga, A. 2001. The OMNeT++ Discrete Event Simulation System. In Proc. of ESM'01. June 6-9, 2001. Prague, Czech Republic

[39] Varga, A. and Hornig, R. (2008), "An overview of the OMNeT++ simulation environment", in Proc. of SIMUTools '08, Marseille, France, March 2008.

[40] 3GPP - TS 36.322 Radio Link Control (RLC) protocol specification

[41] C.Úbeda, S.Pedraza, M.Regueira, J.Romero, "LTE FDD Physical Random Access Channel Dimensioning and Planning", in Proc. of VTC Fall 2012, Sept. 3-6, 2012, Quebec City, CA

[42] William C. Jakes, "Microwave Mobile Communications", John Wiley & Sons Inc, February 1, 1975

[43] IEEE 802.16 Task Group 'm', "IEEE 802.16m Evaluation Methodology Document," Jan. 2009.

[44] ITU-R, M.2135-1, Guidelines for evaluation of radio interface technologies for IMT-Advanced, 12/2009

[45] D. Johnson and D. Maltz. Dynamic source routing in ad hoc wireless networks. In Mobile Computing, pp. 153–181. Kluwer Academic Publishers, 1996.

[46] INET framework for OMNeT++: http://inet.omnetpp.org/ [Accessed December 2013]

[47] IEEE 802.16p, Machine to Machine (M2M) Evaluation Methodology Document, May 24, 2011

[48] T. Tirronen, A. Larmo, J. Sachs, B. Lindoff, N. Wiberg, "Machine-to-machine communication with long-term evolution with reduced device energy consumption", Transactions on Emerging Telecommunications Technologies, 2013

[49] Jha, S.C.; Koc, A.T.; Gupta, M.; Vannithamby, R., "Power Saving mechanisms for M2M communication over LTE networks", in Proc. 1st BlackSeaCom, 2013, pp.102,106, 3-5 July 2013

[50] Girici, T., Zhu, C., Agre, J. R., Ephremides, A. "Proportional fair scheduling algorithm in OFDMA-based wireless systems with QoS constraints", Journal of Communications and Networks, 12(1), 2010, pp. 30-42.

VII. APPENDIX

TABLE 22 – LTE-RELATED ACRONYMS USED IN THE PAPER

Acronym	Definition
BLER	Block Error Rate
BSR	Buffer Status Report
CQI	Channel Quality Indicator
DC	DRX Cycle
DCE	DRX-Command MAC Control Element
DO	DRX Offset
DRX	Discontinuous Reception
DSR	Dedicated Scheduling Request
eNB	Evolved Node-B
H-ARQ	Hybrid Automatic Repeat reQuest
IT	DRX Inactivity Timer
LDC	DRX Long DRX Cycle
LTE	Long-term Evolution
MaxC/I	Maximum Carrier over Interference
ODT	DRX On Duration Timer
PDCCH	Physical Downlink Control CHannel
PDU	Protocol Data Unit
PF	Proportional Fair
RAC	Random Access Procedure
RB	Resource Block
RLC	Radio Link Control
RRC	Radio Resource Control
SCT	DRX Short Cycle Timer
SDC	Short DRX Cycle
SNR	Signal to Noise Ratio
SPS	Semi-Persistent Scheduling
TTI	Transmission Time Interval
UE	User Equipment