Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Reducing Interference via Link Adaptation in Delay-Critical Wireless Networks

Silvio Mandelli (silvio.mandelli@nokia-bell-labs.com)
Nokia Bell Labs https://orcid.org/0000-0002-0166-5544
Alessandro Lieto
Nokia Bell Labs
Mark Razenberg
Nokia Bell Labs
Andreas Weber
Nokia Bell Labs
Thorsten Wild
Nokia Bell Labs

Research

Keywords: Link Adaptation, URLLC, 6G, Power Optimization, Scheduling, Resource Allocation

Posted Date: November 15th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-1034206/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Reducing Interference via Link Adaptation in Delay-Critical Wireless Networks

Silvio Mandelli*, Alessandro Lieto, Mark Razenberg, Andreas Weber and Thorsten Wild

*Correspondence:
silvio.mandelli@Nokia-bell-labs.com
Bell Labs, Nokia, Stuttgart, Germany
Full list of author information is available at the end of the article

Abstract

One of the current 6G wireless networks research's trends is to investigate short distance and dense scenarios, where users are locally connected in sub-networks. Such use case is critical to support the advances of industrial internet of things or Industry 4.0, e.g. connecting an entire group of sensors and actuators of a robot. Therefore, schemes that can properly manage the interference must be deployed in practical systems to allow the promised performance advances of 6G. Targeting these high density scenarios, we describe the Power Optimization for Low Interference and Throughput Enhancement (POLITE) paradigm for link adaptation and power allocation, which leverages available radio resources to stabilize and reduce the interference. The baseline link adaptation schemes are compared with POLITE in their performance in a 3GPP-calibrated system level simulator for industrial scenarios. As services in industrial environments require high reliability under constrained delays, we propose different delay-aware formulations in the POLITE design. In this work we provide solutions both for relaxed delay requirements and for latency critical traffic, whose delay must be minimized. In particular, in the latter case, we propose also modifications of user selection and resource allocation procedures to further improve the reliability and latency. Simulation results prove the benefits of POLITE in terms of increased throughput, fulfillment of relaxed and delay-critical requirements, with an overall reduced transmit power compared to the current baseline link adaptation schemes.

Keywords: Link Adaptation; URLLC; 6G; Power Optimization; Scheduling; Resource Allocation

1 Introduction

With the consolidation of the first releases of 3rd Generation Partnership Project (3GPP) standards for the deployment of the 5-th generation (5G) wireless networks, the research community is already defining the scenarios and directions for 6-th generation (6G) systems [1]. Among the different scenarios envisioned in 6G [2], the 6G short range and low power sub-networks will support life-critical and futuristic applications, like intra-body heart rate control, intra-vehicle break control and Industry 4.0 devices of a central machine hub [3]. Accordingly, the research effort on studying and improving the performance of Ultra Reliable Low Latency Communications (URLLC) is pivotal not only in the legacy macro area wireless systems, but also in these new sub-networks scenarios. It becomes clear the need of simple and efficient distributed algorithms to guarantee co-existence of wireless networks, that would typically operate in an interference-limited regime.

In this work, we address the paradigm for Link Adaptation (LA) and power allocation introduced in [4,5] to improve performance of wireless networks by leveraging all the available radio time-frequency resources of every cell. The Power Optimization for Low Interference and Throughput Enhancement (POLITE) in [4] exploits

the significant underutilization of wireless resources of current systems [6]. By applying POLITE, each transmission can be stretched on a bigger amount of resources, allowing to reduce the transmit power while preserving the transmission's reliability. As a consequence, the system benefits from an overall reduced and stabler interference, that is particularly advantageous for URLLC traffic.

Related work

In the literature, long-term power adaptation techniques are proposed as potential solutions for 6G subnetworks [3], together with uplink power control to reduce interference from close neighbor cell mobiles. However, techniques involving optimization routines [7], game theory [8] or artificial intelligence [9,10] are not tailored for real time Medium Access Control (MAC) procedures. In particular, they either increase significantly the computational complexity, or they require additional signaling among network nodes compared to current Baseline Link Adaptation (BLA) schemes [11] and POLITE. Focusing on recent year's progress in MAC design for URLLC, the two closest approaches to ours are [11,12], where the authors propose to give up in URLLC flows' spectral efficiency to boost reliability. Nonetheless, both approaches do not unveil the potential of more efficient time-frequency resource utilization to reduce and stabilize the interference, that is POLITE's goal. In particular, the better shape of the interference in POLITE brings benefits in terms of higher system capacity and lower overall transmit power compared to the current extensively used BLA paradigm, consisting in the maximization of the user's spectral efficiency.

Although POLITE's benefits are interesting, one of the main limitations of the original POLITE scheme [4] was the unawareness of packet delay budgets and its incapacity to enforce these. The more recent [5] formulates how packet delays can be matched, evaluating its performance in Indoor Factory (InF) scenarios [13], showing performance improvements in terms of system capacity, packet delays fulfillment, and reduced transmit power. However, in the previous works [4, 5], only LA was under the scope of POLITE methods. Focusing on URLLC use cases, in this work we consider how POLITE can interact with user scheduling and resource allocation to minimize latency and increase reliability of latency-critical traffic.

Contribution

The contribution of the paper can be summarized as follows.

- The most recent POLITE schemes [5] have been further enhanced to minimize the latency experienced by critical traffic users, in addition to the mere verification of non-stringent latency budgets. Differently from previous work, we do not only propose a new LA paradigm, but we also provide novel user selection and resource allocation mechanisms. This extension is an enabler for future support of URLLC traffic types.
- Extensive system level simulations of indoor factory floors are reported. Thanks to a proper delay-aware rate reduction policy, it is possible to show more than one degree of magnitude gains in terms of latency critical traffic's reliability compared to baseline schemes, for current 5G QoS Indicator (5QI) [14]. The obtained results validate the POLITE's benefits with short

inter-site distances and the typical severe interference conditions of industrial scenarios, such as the future sub-networks of Beyond 5G and 6G wireless systems.

Structure of the paper

The remainder of the paper is organized as follows. In Section 2, the MAC layer procedures of interest are introduced, namely LA and resource allocation, together with the terminology and the notation adopted in this work. The state of the art LA and the novel POLITE scheme are formally described in Section 3, while in Section 4 and Section 5 the basic and the more advanced delay-aware algorithmic implementations of POLITE are discussed, respectively. Simulation results and numerical experiments are performed in Section 6, where the different approaches are compared with the state of the art. Finally, the summary in Section 7 concludes the paper.

2 MAC layer models

The MAC layer procedures necessary to motivate and understand this paper are (i) LA and (ii) scheduling, consisting in user selection and radio resource allocation. The skeleton of their operation is depicted in Figure 1. At each Transmission Time Interval (TTI) one should i) determine from the set of connected users $\mathcal{C}$, the set of active users $\mathcal{I} \subseteq \mathcal{C}$, that have an amount of bits $Q_i > 0$ to transmit, $\forall i \in \mathcal{I}$, ii) perform LA for each of them and iii) determine which user is scheduled in each Schedulable Resource Block (SRB), defined as the atomic entity of schedulable wireless resources. A SRB can consist of multiple time, frequency and/or spatial radio resource elements. In the next two subsections, we formalize the LA and scheduling models and terminology used in this paper.

2.1 LA Notation

The LA problem in wireless communication can be defined as the selection, for a generic user, of

1. the MCS m, having rate a_m to be used during communication, chosen among the MCS set M, that is sorted with ascending order of rate,
2. the transmit power P - or Power Spectral Density (PSD) S - to be used,

given or subject to

A) maximum transmit power, $P^{(\mathrm{M})}$, or PSD, $S^{(\mathrm{M})}$,
B) an estimate of the channel quality α, that is equal to the equivalent channel power gain divided by the sum of interference plus noise PSD, leading to a Signal to Interference plus Noise Ratio (SINR) $\gamma = S\alpha$,
C) a desired target Block Error Rate (BLER) $\bar{\epsilon}$ for the first transmission attempt,
D) The amount of bits to be transmitted Q.

In MAC procedures, LA is typically performed for every active user and its output is fed to the resource scheduler, that is the object of the next subsection. Note that, assuming to have a pool of $\mathcal{B} = \{b_1, \cdots, b_B\}$ SRBs, one may select a different MCS m_b and transmit power P_b - or PSD S_b - depending on which SRB b is allocated to the user.

2.2 Resource Scheduling

It is important to recall that some resource schedulers, like the well-known round robin scheduling, may not need LA's output before scheduling. However, in wireless literature the most common baseline for scheduling has always been Proportional Fair (PF), thanks to its simplicity and its performance gains on simple baselines. Therefore, we consider a frequency-selective PF scheduler, that computes the following metric

$$M_{i,b} = \frac{a_{m_b^{(i)}}}{r^{(i)}},\ \forall b \in \mathcal{B}, \forall i \in \mathcal{I}, \tag{1}$$

where $m_b^{(i)}$ and $r^{(i)}$ are the MCS assigned by LA to the user i on the SRB b, and user i's temporally *smoothed* past experienced throughput, respectively. The main reason to use PF are its time and frequency opportunistic gains. These are given by the fact that active users obtain in average the same amount of resources, but each user obtains more resources when it can achieve higher spectral efficiency than its average. For more details on PF, or similar scheduling metrics with Quality of Service awareness, the reader can resort to [15].

In this work, unless mentioned otherwise, we consider Algorithm 1, that describes a frequency-selective single-user resource allocation with PF scheduling. It corresponds to the allocation of resources to the combination of user and SRB having the highest PF metric, determined as in (1). The procedure is run until either all resources are exhausted or all active users have emptied their buffer $Q^{(i)}$. After determining which user $W_b \in \mathcal{I}$ is to be scheduled on each SRB $b \in \mathcal{B}$ in the current TTI, this information is passed to the physical layer, together with the user's MCS and PSD, for the effective transmission. If $W_b = 0$, it means that the SRB has not been assigned and nothing is transmitted over those time-frequency resources. The remaining bits to be transmitted for a generic selected user i^* are updated based on the amount of bits that can be transmitted over the overall resources assigned to it in the current TTI, using the selected MCS (see line 10 in Algorithm 1). Note that the MCS decision is updated considering an equivalent SINR over all the resources assigned so far, therefore determining a unique transport block size. More information on the considered MCS, their performance and SINR mappings considered in this work are given in Section 6.

Algorithm 1 PF scheduler basic operations

1: $B \leftarrow \mathcal{B}$ ▷ SRBs to allocate
2: $I \leftarrow \mathcal{I}$ ▷ Active Users
3: $q^{(i)} \leftarrow Q^{(i)}, \forall i \in I$
4: $W_b \leftarrow 0, \forall b \in B$ ▷ User allocated to each SRB
5: Compute $M_{i,b}, \forall b \in B, \forall i \in I$
6: **while** $B \neq \emptyset$ and $\sum_{i \in I} q^{(i)} \neq 0$ **do**
7: $i^*, b^* = \arg\max_{i \in I, b \in B} M_{i,b}$ ▷ User selection
8: $W_{b^*} \leftarrow i^*$
9: $B \leftarrow B \setminus \{b^*\}$
10: Determine $q^{(i^*)}$: remaining bits of user i^*
11: **if** $q^{(i^*)} = 0$ **then**
12: $I \leftarrow I \setminus \{i^*\}$
return $W_b \leftarrow 0$

However, some users may have such tight latency constraints that require more attention. This could be the case of URLLC traffic or any *hard priority* services

that, if active, can completely preempt other users. We can therefore define the set of active hard priority users as $\mathcal{I}' \subseteq \mathcal{I}$. Differently from the pure PF formulation, this information is used by the Proportional Fair with Hard Priority (PFHP) scheduler by changing how the users are prioritized during scheduling. We propose to substitute the user selection procedure of PF (reported line 7 of Algorithm 1) as follows

$$I' = \begin{cases} \mathcal{I} \bigcap \mathcal{I}' & \text{if } \mathcal{I} \bigcap \mathcal{I}' \neq \emptyset \\ \mathcal{I} & \text{otherwise} \end{cases}, \tag{2a}$$

$$i*, b* = \arg \max_{i \in I', b \in B} M_{i,b} \, . \tag{2b}$$

Namely, if there is at least a hard priority user requiring resources, users without hard priority, i.e. in the set $\mathcal{I} \setminus \mathcal{I}'$, are not considered for resource assignment.

3 Link Adaptation Schemes

In this section, we introduce two families of link adaptation algorithms and discuss their problem formulation, by adopting the same formalism introduced in Section 2.

3.1 Baseline Link Adaptation (BLA)

This first algorithm family frames the LA problem as follows

$$m = \sup_m M_{\overline{\epsilon},\gamma,b} = \sup_m \{m \in M : \epsilon_m(\gamma, b) \leq \overline{\epsilon}\} \, , \tag{3}$$

where $\epsilon_m(\gamma, b)$ are the BLER curves for MCS m, the SINR estimate γ and packet size b. Note that the sorting rule of the set of MCS $M_{\overline{\epsilon},\gamma,b} \subseteq M$ satisfying the BLER target is also ordered with ascending rate a_m, hence $m < m' \Leftrightarrow a_m < a_{m'}$, with $m, m' \in M$. This scheme corresponds to the state of the art LA operations, which, therefore, will be referred to as Baseline Link Adaptation (BLA) in what follows. In a more formal way, the BLA approach (3) can also be framed as the rate maximization problem below.

$$\max_m \quad a_m \tag{4}$$

$$\text{s.t.} \quad \epsilon_m(\gamma = S\alpha, b) \leq \overline{\epsilon}, \tag{4a}$$

$$S = S^{(\mathrm{M})}. \tag{4b}$$

3.2 POLITE general concept

In contrast to the BLA approach described in the previous subsection, POLITE exploits additional information to determine the MCS. Previous works [6] demonstrated that typical macro-area wireless systems - working with BLA - are highly under-loaded. Similarly, the more recently addressed 6G sub-networks will likely not operate with full load in every sub-network present in the system if adopting the legacy BLA approach. Therefore, we assume that the MAC layer of each cell can monitor the ratio of resource utilization β if BLA is adopted in the system. If

$\beta < 1$, the system could slow down its transmission rate, allowing to reduce the transmit power accordingly, leading to the POLITE formulation below

$$\min_{m'} \quad S' \tag{5}$$

$$\text{s.t.} \quad \epsilon_{m'}(S'\alpha, b) \leq \bar{\epsilon}\,, \tag{5a}$$

$$a_{m'} \geq \beta a_m\,, \tag{5b}$$

where a_m is the solution of (4). Notice that the same BLER target is guaranteed (cf. Eq. (4a) and Eq. (5a)), but in the problem formulation in (5) the power spectral density is minimized, rather than fixed to a pre-defined value (cf. Eq. (4b)). Note that the POLITE scheme can operate in a distributed way across each single cell, without the need of coordination among them. In [4] the authors proved that POLITE is theoretically able to convey information rates with infinite and finite packet sizes with lower power spectral density and total transmit power. Results in [4] demonstrated that - in urban macro scenarios [13] - reduced transmit power and higher throughput in congested cells can be achieved with POLITE, due to lower and stabler interference from neighboring cells. One of the key challenges of POLITE schemes in practical systems is determining the factor β of Eq. (5b), hereafter referred as slowing factor. Different proposals lead to different effects and will be investigated in the next two sections.

4 Basic POLITE implementation

4.1 Load-driven POLITE (LP)

The first proposal from [4] to estimate the slowing factor β aims at serving all the incoming traffic as follows.

$$\beta^{\text{LP}} = \min\left[\chi \sum_{i\in\mathcal{C}} \frac{\rho_i}{\eta_i N}, 1\right], \tag{6}$$

where we recall that $\mathcal{C}$ is the set of users connected in the considered cell and ρ_i, η_i are, respectively, the exponentially smoothed traffic arrival rate and spectral efficiency of user i, while N the number of resource elements available per TTI. This way to compute β will hereafter be called Load-driven POLITE (LP). Note the insensitivity of LP schemes to packet delays. The multiplicative factor $\chi \leq 1$ considered in (6) is to further enhance the system performance, as shown in [4].

4.2 Aggressive POLITE (AP)

The "Aggressive POLITE" scheme defines an heuristics to handle extreme delay requirements within the POLITE framework. We denote by $\mathcal{C}', \mathcal{I}'$ the sets of (i) connected users and (ii) active users, respectively, with extreme delay requirements. For each of these users, we assume that their transmissions are not affected by POLITE, thus $\beta_j = 1, \forall j \in \mathcal{C}'$. Accordingly, the computation of β for all the remaining users in the set $\mathcal{C}\backslash\mathcal{C}'$, i.e. with no stringent delay requirements, is adapted from the LP formulation in (6) to

$$\beta^{\text{AP}} = \min\left[\chi \frac{\sum_{i\in\mathcal{C}\backslash\mathcal{C}'} \frac{\rho_i}{\eta_i N}}{1 - \min\left[\sum_{j\in\mathcal{C}'} \frac{\rho_j}{\eta_j N}, 1\right]}, 1\right]\,. \tag{7}$$

However, even if it seems counter-intuitive, reducing the rate of urgent (delay-critical) transmissions does not necessarily have a negative impact on their performance. For instance, falling back to more conservative MCS to occupy all remaining unused resources of a single TTI was already suggested in [11] to increase URLLC's reliability. Moreover, if the delay budget allows it, one could also exploit following TTIs for the same purpose [12]. Nevertheless, even if the packet delay budget could be anyway satisfied, previous solutions did not propose to reduce transmit power to reduce and stabilize interference.

5 POLITE with delay budgets

In this section, we account for the latency requirements of critical applications by adopting the specifics of the 3GPP File Transfer Protocol 3 (FTP3) traffic model in [16], where the users in the systems download packets with

- constant size Q,
- exponentially distributed inter-arrival times, with mean μ,
- delay budget - of each packet - B,
- an infinite queue (buffer).

In this work, we assume that a packet is discarded whenever its delay exceeds the packet delay budget. Moreover, all packets of a single user are served according to a First In First Out (FIFO) policy, bringing the system into congestion once the incoming traffic approaches the cell capacity.

5.1 Delay-aware POLITE (DP)

In the LP version [4], the POLITE β is computed considering only an estimate of the arrival rate ρ_i of every connected user in the cell, as done in (6). The limitation of LP is that the estimate of the required amount of resources does not account for delay-sensitive transmissions, whose requests must be served within a certain delay budget. Therefore, Delay-aware POLITE (DP) implements a more general rule to estimate the β to be used by POLITE when performing LA. In particular, for every packet $p \in P_i$ waiting in the i-th user's queue, we consider

- the amount of data still to be transmitted $Q_{i,p}$,
- the arrival time $T_{i,p}$,
- an arbitrary/configurable "desired" delay $D_{i,p} \leq B_{i,p}$,
- an arbitrary/configurable "delay normalizing factor" $D^0_{i,p} \geq 0$, with $D^0_{i,p} \ll B_{i,p}$.

The scheduler can then estimate the i-th user's required throughput at time t, $\rho'_i(t)$, by taking the maximum value between the long-term arrival rate and the maximum throughput needed to serve all its packets within their latency budgets:

$$\rho'_i(t) = \max \left[\chi \hat{\rho}_i(t), \max_p \left[1/\tau_{i,p}(t) \sum_{k=0}^{p} Q_{i,k} \right] \right], \tag{8}$$

where $\hat{\rho}_i(t)$ is the single user exponentially smoothed average arrival rate at time t, as done in [4]. Therefore, the DP's β is computed as

$$\beta^{\mathrm{DP}} = \min \left[\sum_{i \in \mathcal{C}} \frac{\rho'_i}{\eta_i N}, 1 \right]. \tag{9}$$

Note that the factor χ is applied only to the long-term average incoming traffic and not the required throughput to serve packets within their delay budgets. For the remainder of the subsection, the dependency on the user i is removed for ease of notation. The value $\tau_p(t)$ (i.e., $\tau_{i,p}(t)$) in Eq. (8) defines the throughput needed to serve a packet p within delay budget,

$$\tau_p(t) = \min\left[\tau_p^{\text{act}}(t), \max\left[D_p^0, \tau_p^{\text{des}}(t)\right]\right], \tag{10}$$

where

- $\tau_p^{\text{act}}(t)$ denotes the actual expiring time, namely $\tau_p^{\text{act}}(t) = \max\left[\tau^{\min}, B_p - t + T_p\right]$, and
- $\tau_p^{\text{des}}(t)$ is an *artificial* desired expiring time, i.e. $\tau_p^{\text{des}}(t) = \max\left[\tau^{\min}, D_p - t + T_p\right]$.

An illustrative example of the quantity $\tau_p(t)$ in Equation (10) has been plot in Figure 2 for the parameters $B_p = 30$ ms, $D_p = 15$ ms, $D_p^0 = 5$ ms and $\tau^{\min} = 1$ ms. In particular, the choice of $\tau^{\min} = 1$ ms reflects the scheduling TTI duration and it is introduced for practical implemention to avoid divisions by zero in (8). From the graphical representation, one can observe the effect of each component defined in (10). The term $\tau_p^{\text{act}}(t)$ is a linear decreasing function of the delay budget and elapsed time (dotted purple line with circles). Considering only such a value, DP would not react in time if deep channel fades occur at the end of the delay budget, making some packets fail. To avoid that the system accumulates packet close to their expiring time B_p, the desired delay $D_p < B_p$ is introduced in $\tau_p^{\text{des}}(t)$. Accordingly, e.g. with $D_p = 15$ ms, the system reacts faster to deliver the packets. Nevertheless, big fluctuations of the β^{DP} may occur due to the unnecessary low values of $\tau_p^{\text{des}}(t)$ when approaching the desired delay. Therefore, a flat region, corresponding to $D_p^0 = 5$ ms, is interposed between the two lines between the desired and the target delay budget.

5.2 Latency-Critical POLITE (LCP)

Algorithm 2 LCP algorithm for Resource Scheduling and Link Adaptation

Perform LA and RS with DP and PFHP;
if not all $p \in P'$ will be fully transmitted **then**
  $B \leftarrow \mathcal{B}$ ▷ SRBs to be allocated
  $\hat{I} \leftarrow \mathcal{I}'$ ▷ Active HP Users
  $q^{(i)} \leftarrow Q^{(i)},\ \forall i \in \hat{I}$ ▷ Bits to be transmitted
  Compute $M_{i,b}\ \ \forall b \in B,\ \forall i \in \hat{I}$ ▷ MCS from BLA
  $\mathbf{R} \leftarrow \text{PFHP}(B, \hat{I}, q^{(i)})$ ▷ Allocation vector for HP users from Algorithm 1
  Initialize null vectors $\mathbf{W}, \mathbf{R}'$
  while $\sum_{i\in\hat{I}} R_i + R'_i \leq |B|$ **do** ▷ Start WRR
    $i^* = \{i^* : W_{(i^*)} = \min(W)\}$ ▷ User selection
    $R'_{(i^*)} \leftarrow R_{(i^*)} + 1$ ▷ Increase allocation
    $W_{(i^*)} \leftarrow W_{(i^*)} + \frac{1}{R_{(i^*)}}$ ▷ Update WRR weights
  while $B \neq \emptyset$ **do**
    $i^*, b^* = \arg\max_{i\in\hat{I}, b\in B} M_{i,b}$
    $W_{b^*} \leftarrow i^*$
    $B \leftarrow B \setminus \{b^*\}$
    if $|\{b \in B | W_b = i^*\}| = R_{(i^*)} + R'_{(i^*)}$ **then**
      $\hat{I} \leftarrow \hat{I} \setminus \{i^*\}$
  Set β_i^{LCP} as in Eq. (11)
  Perform LA using β_i^{LCP}

Although DP allows to satisfy delay budgets, it does not aim at minimizing the delay for latency-critical users. For some type of traffic, we do not only need to meet their latency budgets, but also to minimize the overall latency. For example, latency-critical applications like URLLC are very sensitive to packet delay, and they should be transmitted in the shortest time possible. By using DP with $\beta^{DP} < 1$, it could occur that the transmission of some latency-critical packets cannot be completed at the current TTI, while with BLA it could have been possible to fully transmit it. As a result, these packets will experience an increase in delay with DP, with a negative impact on their performance. As a solution one might consider to apply AP, thus no rate reduction of latency-critical traffic. However, this leads to interference spikes that further lower performance, as demonstrated in [5] and will be discussed in Section 6. Therefore, we introduce the Latency-Critical POLITE (LCP), an extension which aims to minimize the delay of latency-critical packets, while still achieving power reduction.

The LCP algorithm works as follows: first, it applies DP and Resource Scheduling (RS) with PFHP to assign to the latency-critical users the highest priority (line 1 of Algorithm 2). If all the latency-critical packets are transmitted at the current TTI, there is no need to intervene, the latency is minimized and DP can be executed. If there are still some pending transmissions for latency-critical users, the following procedures is considered. With BLA and PFHP the initial resource allocation for delay-critical users is computed (lines 6-7 of Algorithm 2). If there are still available resources, these are further distributed among the delay-critical users in a Weighted Round Robin (WRR) fashion (see lines 9-12 of Algorithm 2). Notice that, to account for the discrete nature of SRBs, WRR allocates resource blocks one per time (line 11), until completion of resources, which might result in an uneven allocation of resources among users. The actual allocation of resources is then computed with a redefined version of PFHP (lines 13-18 in Algorithm 2), where line 17 defines the maximum resource allowance of a user, as defined by WRR in the previous steps of the algorithm. Finally, based on the amount of resources to be assigned to each user, a per-user slowing factor β can be calculated as

$$\beta_i^{\mathrm{LCP}} = \min\left(1, \frac{\sum_{k=0}^{p} Q_{i,k}}{\xi_i}\right) \tag{11}$$

where the denominator ξ_i is the amount of bits can be transmitted to user i when using the MCS determined with BLA on the SRBs allocated to i. Finally, LA is performed on the scheduled SRBs, using for every user its corresponding β_i^{LCP} (line 20 of Algorithm 2).

The reasoning behind the computation of BLA and PFHP in line 6-7 of Algorithm 2 is that it ensures the transmission of all latency-critical packets, if enough resources are available. Only after the transmission of those packets is guaranteed, additional SRBs are exploited to reduce the required Tx power. In particular, differently from DP, LCP estimates a per-user slowing factor β_i, by allowing different power reduction factors for different latency-critical users. This choice is needed due to the integer - and not rational - number of SRBs allocated to each user. Therefore, one could not arbitrarily redistribute resources to have a common slowing factor. However, one should rather compute a per-user slowing factor depending on each user's buffer size and resource elements allocated to it in the current TTI.

6 Simulation Method, Results and Discussion

6.1 Simulation method and scenario description

The experiments are performed in a downlink (DL) system-level simulator, since implementing the proposed scheme in DL does not require any additional signaling between base station and users [4]. The 3GPP calibrated InF [13] channel model is considered in our simulations, abstracting the physical-layer effects through a link to system-level interface computing equivalent SINR at transmission time, given the cell/user topology and active transmissions. The simulation environment consists of a confined industrial area of 120x60 m^2, split into a 6x3 cell topology with 20 m inter-site distance. The main simulation parameters can be found in Table 1.

General Environment	3GPP Dense High (DH) InF [13]
Cell deployment	6x3, inter-site distance 20 m
Simulation experiments	20 experiments of 25 s
FTP3 Users per Cell	FTP3-L(ong) 10, FTP3-S(hort) 20
Full Buffer Users in total	1
Interference Estimation	Data based with 100 ms avg.
Channel Quality Indicator (CQI)	Sub-band reports every 5 ms
CQI/MCS Table	CQI/MCS Table 3 [17]
Link Performance	Low-Density Parity Check (LDPC) data channel codes [17, 18]
Subcarrier Spacing, TTI length	15 kHz, 1 ms
Central Frequency, Bandwidth	3.5 GHz, 10 MHz
Max Transmit Power	30 dBm
Base Station Antennas	1 (or 2) Vertically polarized
User Antennas	1 (or 2) Vertically polarized
Spatial multiplexing	1(or 1-2) layers Single User MIMO
Receive combining	IRC [19]
User Mobility	3 km/h
Link to System-Level Model	[20]

Table 1: Baseline simulation parameters

The interference estimated by each user (in terms of receive antenna co-variance matrix) is performed by observing real data transmission, thus knowing if there were interfering data transmissions in the past. The maximum transmit power is enforced with a power spectral density limitation over all the available subcarriers at every TTI. Therefore, the total irradiated power is lower if some subcarriers are not used for transmission. In case of 2x2 MIMO simulations, an Interference Rejection Combining (IRC) receiver [19] is used to compute the receive combining weights and the transmission rank is determined by the base station based on the maximum throughput achievable. The traffic is mainly due to two classes of FTP3 [16] users randomly deployed in the system, that download packets of fixed dimensions with varying packet exponential inter-arrival time, allowing to tune the average offered load in each cell. The two classes FTP3 Long (FTP3-L) and FTP3 Short (FTP3-S) represent long, but relaxed, file transfers and short, but delay-critical, packet transmissions, respectively, and their parameters are resumed in Table 2. The delay-critical traffic users have a latency target of 30 ms, that is corresponding to the delay of “Intelligent Transport System” 5QI number 84 of “Delay-Aware GBR” services in 5G systems [14]. Note that enabling sub-ms delays for future 6G sub-networks applications could be achieved by considering systems with shorter TTIs than 1 ms, as shown in [11]. Compared to the simulations in [5], the FTP3-S packet dimension and arrival rate has been increased to challenge the system with bigger and more frequent bursts of delay critical traffic. Moreover, differently from [5], we always

assign hard priority at scheduling to FTP3-S traffic to reduce its delay as much as possible. The incoming traffic estimation by every user is performed observing the buffer status changes and applying exponential smoothing, more details are given in [4]. Finally, a Full Buffer (FB) user having always data to transmit is placed in a random cell. In this cell, all resources will be always occupied due to this user, that has an infinite thirst for resources.

Parameter	FTP3-L	FTP3-S
Average number of users per cell	10	10
Packet dimension (Q)	800 kbits	2048 bits
Packet arrival rate [pkt/s/user]	λ_L	10
Delay Budget (B) [ms]	4000	30
Desired Delay (D) [ms]	3000	15
Delay Normalizer (D_0) [ms]	250	5

Table 2: FTP3 traffic parameters

The analyzed algorithms are

- the original LP of Subsection 4.1 with $\chi = 0.8$.
- the delay-aware DP of Subsection 5.1 with $\chi = 0.8$.
- the more aggressive AP of Subsection 4.2, but only applied for FTP3-S traffic, while FTP3-L delay is still enforced with DP with $\chi = 0.8$.
- the final LCP proposal of Subsection 5.2 with $\chi = 0.8$. As in the previous case, LCP is only applied for FTP3-S traffic, while FTP3-L delay is still enforced with DP.
- The state of the art BLA of Subsection 3.1.

The rationale behind the adoption of χ is that the finite set of available MCS (i) limits the minimum available rate and (ii) forces the MCS selection of Equation (5b) to have a rate typically higher than the desired one. Moreover, (iii) the channel selectivity in frequency and the opportunistic nature of the PF scheduling described in Algorithm 1 allow users to be allocated resources where their rate can be higher than the average rate over the full bandwidth. Therefore, a further slowing factor $\chi < 1$ can push the resource utilization closer to 100%, thus increasing POLITE's performance [4].

In this work, the discussion in [4, 5] is extended by analyzing (i) LCP's performance, and (ii) the tradeoff between the delay budgets satisfaction and generic system performance of different POLITE algorithms. To improve clarity, the result section is structured in two main parts. In the first, packet delay performance is addressed, whereas in the second part generic system performance are investigated, namely the FB throughput and the required average transmit power.

6.2 Delay results

In order to assess the ability of the proposed schemes to convey the offered traffic within its delay budget, the performance of the two traffic types, FTP3-S and FTP3-L, are discussed separately. The performance are evaluated in terms of average and, for FTP3-S, distribution of the transmission delay. Moreover, we analyze the failure ratio, measured as the ratio between the number of packets that cannot be transmitted within their delay budget versus the total number of offered packets in the system. In the following plots, the same colors (and markers) are used to distinguish the different algorithm proposals.

FTP3-S average delay and failure ratio In Figure 3, the average delay and the packet failure ratio for the FTP3-S are shown, respectively in solid lines and dotted lines. One can see that, as general trend, the performance of the FTP3-S users worsen as the offered traffic of FTP3-L users increases, due to the corresponding interference generated, reducing the spectral efficiency, thus capacity, of neighboring cells. In particular, one can appreciate that the average delay is minimized by the proposed LCP (solid purple line), followed by AP, which consists of the most conservative approach to serve the delay-critical users. Note that the interference spikes of AP still make it underperform compared to LCP. On the contrary, especially at low load, the LP and DP exhibit higher average delay, due to the fact that they do not aim at minimizing latency. The failure ratio plot in Figure 3 with dashed lines allows to analyze how the investigated algorithms match delay budgets: the delay un-aware LP, as expected, underperforms compared to BLA at low offered traffic due to its excessive tendency of slowing down, improving at higher loads. Then, both the AP and the DP can match FTP3-S delay budgets better than BLA at all loads. However, the LCP is the only one that manages to significantly improves the performance of the FTP3-S traffic for any cell load condition, achieving more than one order of magnitude less failure ratio than its competitors at low loads, with its gains shrinking, but not nulling, as the load increases.

FTP3-S Complementary Cumulative Density Function (CCDF) To better describe the behavior of the LA schemes in different load conditions, in Figure 4 the CCDF of the packet delay is plotted in two scenarios: a high and a low interference scenario, consisting of 0.4 and 4 Mbps of FTP3-L offered traffic per cell. One can notice the impact of the interference from other cells in the drastic increase of the latency, thus of the probability of experiencing latencies of 30 ms or more, thus failing the delay budget. Except from LP at low loads, it's worth noticing the better behavior of all POLITE proposals compared to BLA regarding failure ratio. In particular, at low load (0.4 Mbps of FTP3-L traffic) DP tries only to match the delay budget, achieving higher latencies than BLA, whereas LP is completely outperformed.On the other hand, the two approaches that try to minimize latency, i.e. AP and LCP, achieve always better performance than BLA, with the latter achieving lower delay than the former. In particular, LCP outperforms BLA by one order of magnitude in CCDF already after 8 ms delay.

In Figure 5 the performance of the FTP3-L users are shown, in terms of packet average delay and failure ratio. One can observe that all POLITE proposals, due to their slowing down of the FTP3-L best effort traffic, experience higher average delay, especially at low load, compared to BLA. Moreover, due to lower amount of resources needed for FTP3-S of AP and LCP, their average delay is lower at low to mid loads compared to DP and LP. However, as the load increases the better capacity of DP to handle interference and loose delay budgets makes it the best POLITE proposal in terms of average delay, with less than 100 ms losses with respect to BLA, over a 3 seconds delay budget. While with POLITE all the transmissions are slowed down, resulting in higher average delay, if one analyses the FTP3-L failure ratio, the trend changes. Given the reduced and stabler interference

compared to BLA, the network performance can increase (which will be addressed further in the next subsection), and, thus, even the failure ratio of low priority traffic can be reduced. By analyzing the failure ratio (dashed lines) in Figure 5, one can notice that BLA, although not slowing down any user, provides a worse performance than all the POLITE schemes, with the exception of the delay unaware LP. In particular, the FTP3-L load, at which the failure ratio is greater than 1%, is ca. 1.1, 1.45, 1.45 and 1.7 Mbps for BLA, AP, LCP and DP, respectively. This shows that AP, LCP and DP can bear ca. 32%, 32% and 54% higher FTP3-L load than BLA if the failure rate target is 1%. One should notice that at high loads, BLA slightly improves performance compared to LCP - applied for FTP3-S only - showing its negative impact on FTP3-L failure ratio. However, this shortcoming is highly compensated by the gains achieved by LCP on FTP3-S's delay (as shown in Figure 3 and Figure 4). On the other hand, the DP solution already introduced in [5] remains the best option if the served traffic does not require to minimize its latency, outperforming all the other proposals in terms of failure ratio, at any load condition, and average delay at mid-high loads.

6.3 FB Throughput and Tx Power comments

In this subsection the overall system performance is evaluated and compared among the baseline BLA and the diverse POLITE variants. In Figures 6 and 7, the achieved throughput of the FB user is plotted, allowing to assess the performance in cells that are fully congested, while others operate with varying load conditions. This situation may happen in realistic systems, for instance in industrial scenarios where big downloads may happen on top of normal traffic, e.g. due to log data or firmware downloads, requiring full capacity in certain cells. The results in Figure 6 show that the proposed POLITE schemes achieve higher throughput than BLA, when considering a 1x1 antenna configuration. In particular, when comparing the POLITE schemes amongst themselves, one can notice that LP outperforms the delay-aware schemes (both LCP and DP). This is the price to pay in order to meet the delay targets of FTP3-L and FTP3-S users, which significantly improve their performance, especially the FTP3-S thanks to LCP. In particular, they can improve performance, by providing ca. a boost of $80-90\%$ of achieved throughput at low-mid loads compared to BLA. Note that, as already stated throughout the paper, AP performs as the worst among the POLITE schemes, due to the spikes of interference generated by the transmission of delay-critical users. As expected, the gain in performance starts decreasing when the load of FTP3-L users increase, almost vanishing when the system operates at almost full capacity. This result can be explained by the difficulty of the POLITE schemes to further reduce transmission rates in congested scenarios. Still small gains can be achieved, thanks to the ability of POLITE to reduce interference when single cells are not fully loaded.

The gains are more evident in the 2x2 MIMO configuration, in Figure 7, where one can notice an improvement of ca. 100% in the region between 2 Mbps and 5 Mbps of carried load. Also for the 2x2 MIMO scheme, the same observations can be drawn when comparing the different POLITE proposals. Interestingly, the gains are higher at mid loads than at high loads - as with 1x1 MIMO - and low loads. At particularly low loads, 2x2 MIMO allows to deliver all the offered FTP3-S and

FTP3-L traffic with minimal resource consumption. Therefore, there could be many transmissions without any interference from the neighboring cells with BLA, thus not requiring the interference reduction properties of the POLITE schemes.

Similar considerations can be done when looking at the plots in Figure 8, where the average transmitted power is plotted for all the tested LA schemes. When comparing the different investigated solutions, transmit power savings can be sorted as LP > DP > LCP > AP $\gg$ BLA. The gains of DP and LCP schemes are more significant at mid low, with a reduction of ca. $4-5$ dB at around 1 Mbps in 1x1 configuration (solid line) and $5-6$ dB at around 5 Mbps for the 2x2 MIMO configuration (dashed line). However, one can notice that, in case of 2x2 MIMO and at very low loads, BLA requires a lower average transmit power. This is due to the fact that at extremely low loads, transmitting short bursts of traffic could be done with a higher energy efficiency due to a negligible probability of being interfered. However, the rare interference events at full power of BLA still provide lower performance in terms of both throughput for full buffer users (Figures 6, 7) and latency for delay-critical users (Figures 3, 4).

Concluding, we highlight that LCP's gains of URLLC performance observed in the previous subsection is traded-off for minor losses in terms of peak throughput and overall needed transmit power compared to DP. Still, the overall performance of DP and LCP proposals severely outperform the baseline scheme BLA, justifying them as suggested schemes to be adopted in future wireless networks.

7 Conclusion

The proposed POLITE methods leverage the unused wireless time-frequency resources in non-congested cells to minimize the transmit power, thus interference, allowing overall performance gains in the system. In this work, delay of latency critical traffic has been minimized by introducing the Latency-Critical POLITE (LCP) paradigm, that integrates modifications of current link adaptation, user scheduling and resource allocation procedures. System level simulations of 3GPP indoor factory scenarios show that, compared to state of the art mechanisms, LCP can improve reliability up to an order of magnitude compared to state of the art BLA approaches, outperforming also previous POLITE proposals.

The important gains for latency critical traffic require a minor price to pay in terms of overall system performance. Still, the LCP proposal delivers from circa 60% to 100% more bearable throughput for full buffer users in the congested cell, more than 4 dB transmit power reduction in non congested cells, 32% increased bearable load with 1% failures of traffic with long deadlines, when it is compared to baseline BLA, in the considered scenario. Therefore, the proposed POLITE paradigm is particularly promising due to its simple and distributed implementation, that leads to improved system performance for all the traffic types that are of interest in current and future wireless systems.

Acronyms

3GPP 3rd Generation Partnership Project.

5QI 5G QoS Indicator.

AP Aggressive POLITE.

BLA Baseline Link Adaptation.
BLER Block Error Rate.

CCDF Complementary Cumulative Density Function.
CQI Channel Quality Indicator.

DP Delay-aware POLITE.

FB Full Buffer.
FIFO First In First Out.
FTP3 File Transfer Protocol 3.
FTP3-S FTP3 Short.
FTP3-L FTP3 Long.

InF Indoor Factory.
IRC Interference Rejection Combining.

LA Link Adaptation.
LCP Latency-Critical POLITE.
LP Load-driven POLITE.

MAC Medium Access Control.
MCS Modulation and Coding Scheme.

PF Proportional Fair.
PFHP Proportional Fair with Hard Priority.
POLITE Power Optimization for Low Interference and Throughput Enhancement.
PSD Power Spectral Density.

RS Resource Scheduling.

SINR Signal to Interference plus Noise Ratio.
SRB Schedulable Resource Block.

TTI Transmission Time Interval.

URLLC Ultra Reliable Low Latency Communications.

WRR Weighted Round Robin.

Declarations

Availability of data and materials
Data can be replicated with 3GPP calibrated system level simulators, according to [13], with the scenario described in Section 6 of this document.

Competing interests
Not applicable.

Funding
Not applicable.

Authors' contributions
Silvio Mandelli and Alessandro Lieto have the main paper contribution on the ideas, simulation study and writing. Mark Razenberg contributed to the development and implementation of the most recent contribution, the LCP algorithm. Andreas Weber and Thorsten Wild supported the work with ideas and precious feedback.

Acknowledgements
The authors thank Paolo Baracca for the fruitful exchange when developing the initial POLITE concept and his support.

Authors' information
If needed, they will be added in the camera ready version.

Author details
Bell Labs, Nokia, Stuttgart, Germany.

References

1. Zhang, Z., Xiao, Y., Ma, Z., Xiao, M., Ding, Z., Lei, X., Karagiannidis, G.K., Fan, P.: 6G wireless networks: Vision, requirements, architecture, and key technologies. IEEE Vehicular Technology Magazine **14**(3), 28–41 (2019)
2. Viswanathan, H., Mogensen, P.E.: Communications in the 6G Era. IEEE Access **8**, 57063–57074 (2020)
3. Berardinelli, G., Baracca, P., Adeogun, R.O., Khosravirad, S.R., Schaich, F., Upadhya, K., Li, D., Tao, T., Viswanathan, H., Mogensen, P.: Extreme communication in 6g: Vision and challenges for 'in-x' subnetworks. IEEE Open Journal of the Communications Society, 1–1 (2021). doi:10.1109/OJCOMS.2021.3121530
4. Mandelli, S., Lieto, A., Baracca, P., Weber, A., Wild, T.: Power optimization for low interference and throughput enhancement for 5G and 6G systems. In: 2021 IEEE Wireless Communications and Networking Conference Workshops (WCNCW) (IEEE WCNC 2021 Workshops), Nanjing, China (2021)
5. Mandelli, S., Lieto, A., Weber, A., Wild, T.: Power optimization and throughput enhancement in 6g networks by delay-aware resource leverage. In: 2021 Joint European Conference on Networks and Communications & 6G Summit (EuCNC/6G Summit), pp. 176–181 (2021). IEEE
6. Park, S., Agiwal, M., Kwon, H., Jin, H.: An evaluation methodology for spectrum usage in LTE-A networks: Traffic volume and resource utilization perspective. IEEE Access **7**, 67863–67873 (2019)
7. Ramezani-Kebrya, A., Dong, M., Liang, B., Boudreau, G., Seyedmehdi, S.H.: Joint power optimization for device-to-device communication in cellular networks with interference control. IEEE Transactions on Wireless Communications **16**(8), 5131–5146 (2017)
8. Miao, G., Himayat, N., Li, G.Y., Talwar, S.: Distributed interference-aware energy-efficient power optimization. IEEE Transactions on Wireless Communications **10**(4), 1323–1333 (2011)
9. Wijaya, M.A., Fukawa, K., Suzuki, H.: Intercell-interference cancellation and neural network transmit power optimization for MIMO channels. In: VTC2015-Fall (2015). IEEE
10. Adeogun, R.O., Berardinelli, G., Mogensen, P.E.: Learning to dynamically allocate radio resources in mobile 6g in-x subnetworks. In: IEEE International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC) (2021)
11. Pocovi, G., Pedersen, K.I., Mogensen, P.: Joint link adaptation and scheduling for 5G ultra-reliable low-latency communications. IEEE Access **6**, 28912–28922 (2018)
12. Belogaev, A., Khorov, E., Krasilov, A., Shmelkin, D., Tang, S.: Conservative link adaptation for ultra reliable low latency communications. In: IEEE International Black Sea Conference on Communications and Networking (BlackSeaCom) (2019)
13. 3GPP: TR 38.901. Study on channel model for frequencies from 0.5 to 100 GHz. Technical report (2020)
14. 3GPP: TS23.501. System architecture for the 5G System (5GS). Technical specification (2020)
15. Mandelli, S., Andrews, M., Borst, S., Klein, S.: Satisfying Network Slicing Constraints via 5G MAC Scheduling. In: IEEE Conference on Computer Communications (INFOCOM) (2019)
16. 3GPP: TR 36.872. Small cell enhancements for E-UTRA and E-UTRAN - Physical layer aspects. Technical report (2013)
17. 3GPP: TS38.214. NR; Physical layer procedures for data. Technical specification (2020)
18. 3GPP: TS38.212. NR; Multiplexing and channel coding. Technical specification (2020)
19. 3GPP: Study on Network-Assisted Interference Cancellation and Suppression (NAIC) for LTE. Technical report (2014)
20. Wan, L., Tsai, S., Almgren, M.: A fading-insensitive performance metric for a unified link quality model. In: IEEE Wireless Communications and Networking Conference, WCNC 2006.

Figures

The figures' title and legend are included in each figure separately, in agreement with the committee of reviewers.

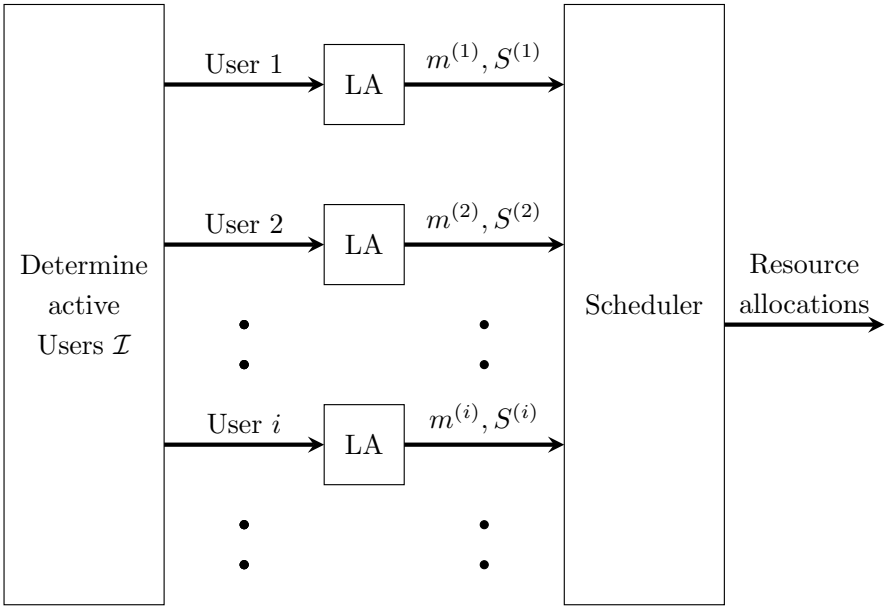

Figure 1: Block diagram of MAC layer operations to handle all users active in a given TTI. The terms $m^{(i)}$ and $S^{(i)}$ are the selected Modulation and Coding Scheme (MCS) and transmit power spectral density for the User i, respectively.

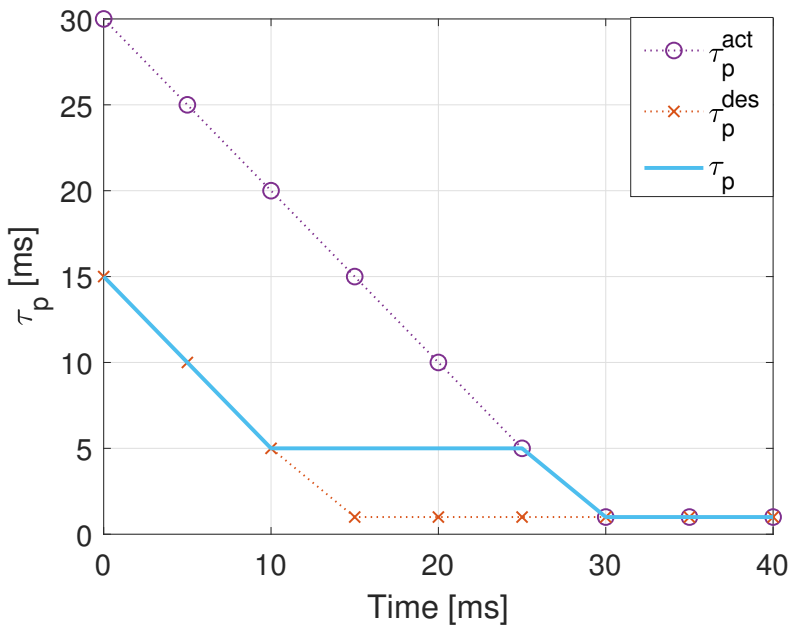

Figure 2: Desired remaining delay curve used for DP.

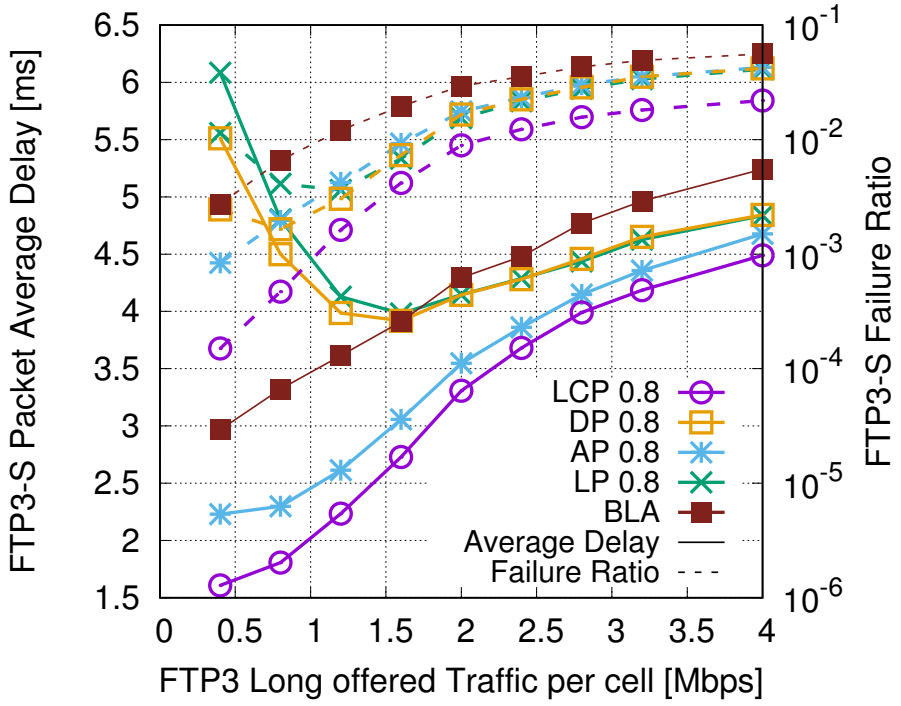

Figure 3: FTP3-S failure ratio and average delay with 1x1 MIMO.

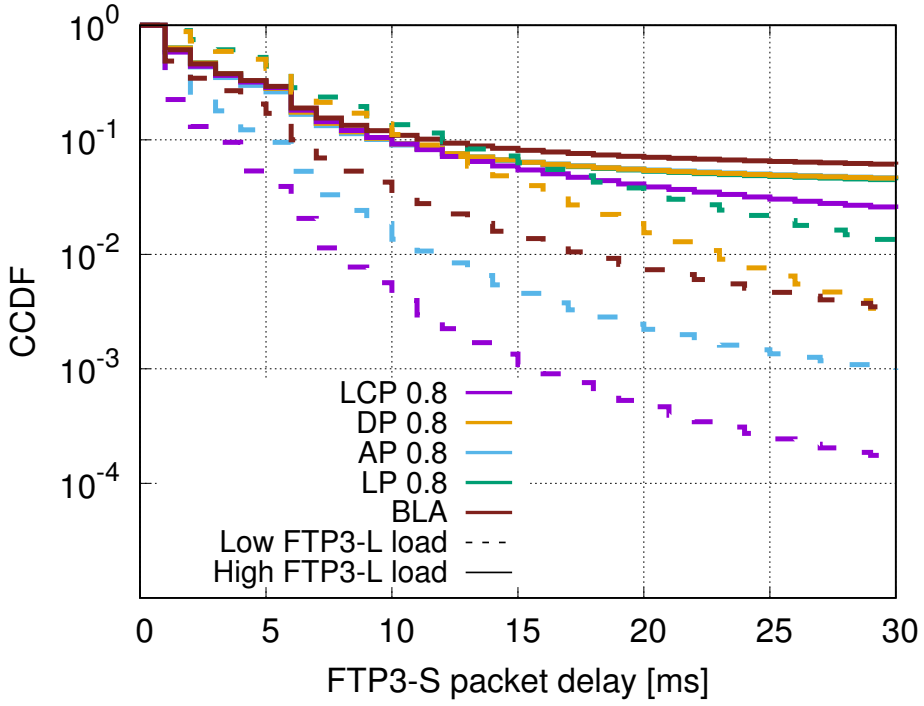

Figure 4: CCDF of delay of FTP3-S traffic under 0.4 and 4.0 Mbps FTP3-L offered traffic per cell with 1x1 MIMO.

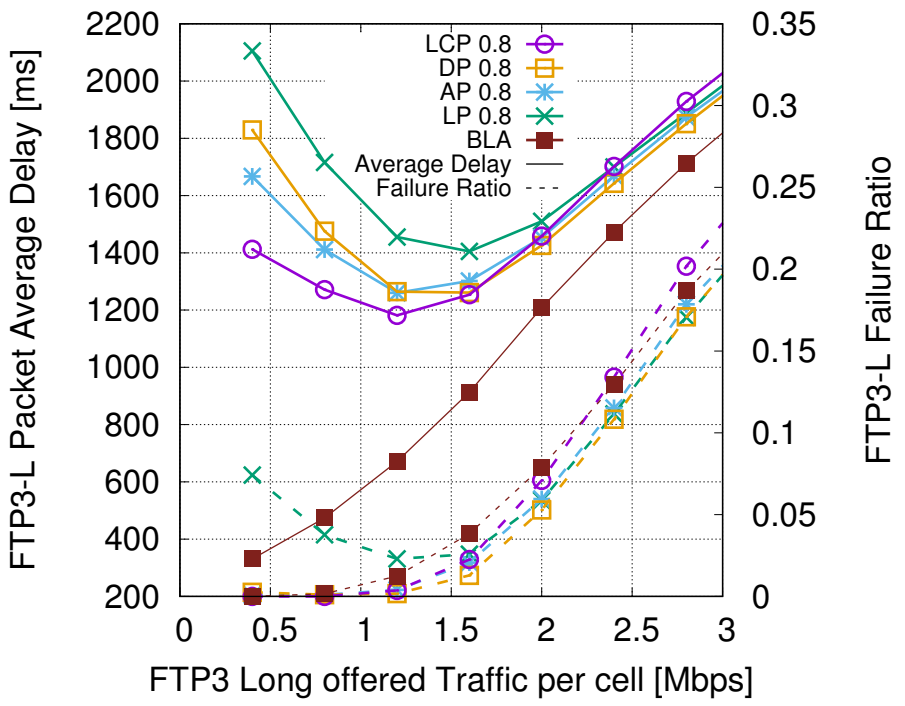

Figure 5: FTP3-L packet deadline failure ratio and average delay with 1x1 MIMO.

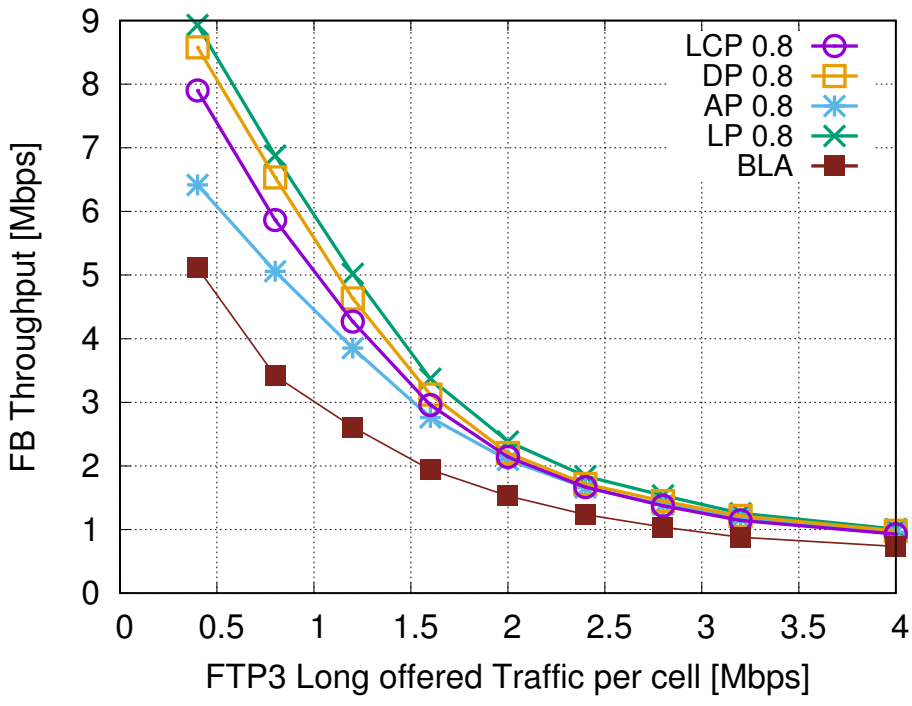

Figure 6: Average FB thoughput in cells with 1x1 MIMO

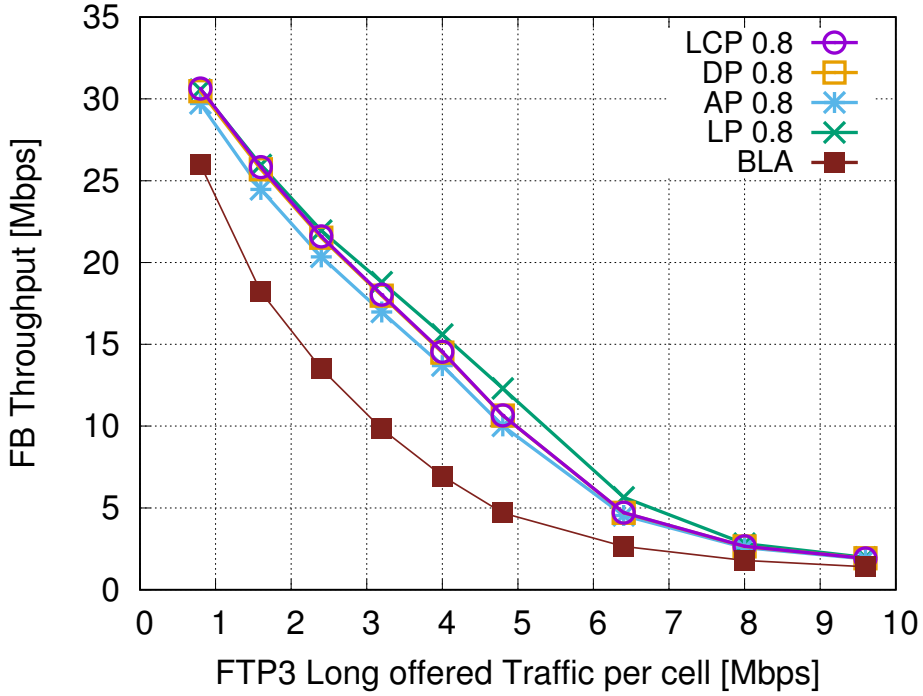

Figure 7: Average FB thoughput in cells with 2x2 MIMO

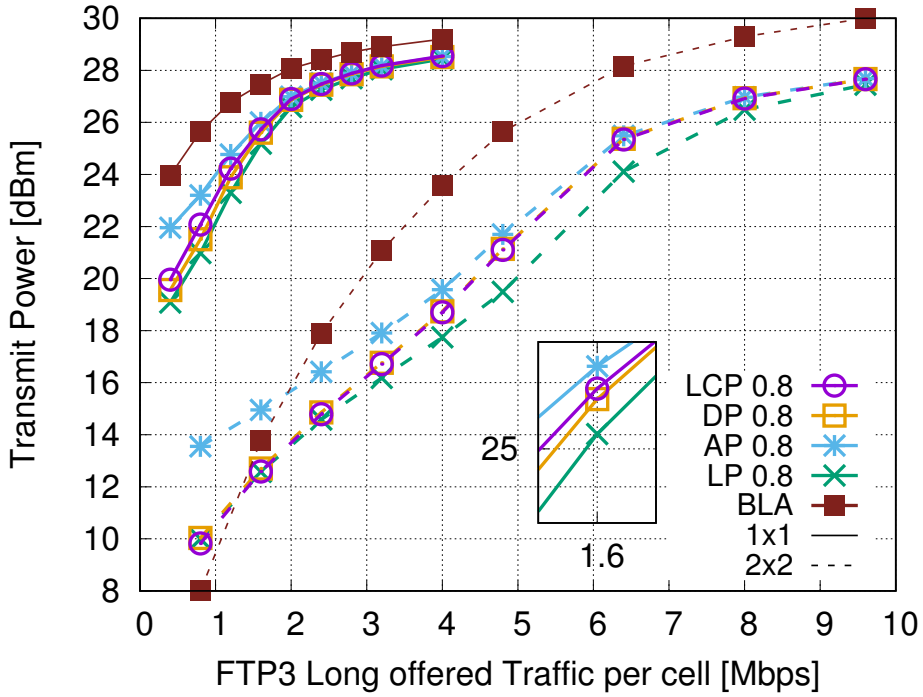

Figure 8: Average transmit power in cells with only FTP3 traffic with 1x1 and 2x2 MIMO