Outage Analysis of Downlink URLLC in Massive MIMO systems with Power Allocation

Alexandru-Sabin Bana1, Luca Sanguinetti2, Elisabeth De Carvalho1, and Petar Popovski1

1Department of Electronic Systems, Aalborg University, Denmark
2Dipartimento di Ingegneria dell’Informazione, University of Pisa, 56122 Pisa, Italy
Email: 1{asb, edc, petarp}@es.aau.dk, 2luca.sanguinetti@unipi.it

Abstract—Massive MIMO is seen as a main enabler for low-latency communications, thanks to its high spatial degrees of freedom. The channel hardening and favorable propagation properties of Massive MIMO are particularly important for multiplexing several URLLC devices. However, the actual utility of channel hardening and spatial multiplexing is dependent critically on the accuracy of channel knowledge. When several low-latency devices are multiplexed, the cost for acquiring accurate knowledge becomes critical, and it is not evident how many devices can be served with a latency-reliability requirement and how many pilot symbols should be allocated. This paper investigates the trade-off between achieving high spectral efficiency and high reliability in the downlink, by employing various power allocation strategies, for maximum ratio and minimum mean square error precoders. The results show that using max-min SINR power allocation achieves the best reliability, at the expense of lower sum spectral efficiency.

I. INTRODUCTION

A massive MIMO system is comprised of a base-station (BS) equipped with a very large number of antennas $M$, which are controlled in a fully digital manner, in order to spatially multiplex a large number of devices simultaneously [1]. The large number of antennas at the BS provide a tremendous increase in the number of degrees-of-freedom (DoFs), which is paramount in achieving the design goals of 5G [2], [3], [4], [5].

Massive MIMO is largely considered as the main technology enabling the high data-rates and high spectral efficiency (SE) required by the extended mobile broadband (eMBB) service within 5G [1], [4], [2], [6]. However, together with the roll-out of 5G, new services have emerged, complementary to eMBB. These services are tailored to serving machine-type communications (MTC) and, whether we are regarding massive MTC (mMTC) or ultra-reliable low-latency communications (URLLC), massive MIMO is considered to be their key enabler [5] as well.

The requirements of URLLC are considerably stricter than previous services. According to 3GPP [7], the typical reliability of a URLLC packet of 32 bytes is defined to be 99.9999% within 1 ms latency. The challenges and potential solutions for achieving the strict URLLC requirements have been largely discussed in the literature [8], [9], [10], [5], and it is worth mentioning a few notable enablers such as: interface diversity[11], network slicing [10], forward error correction (FEC) for code diversity [12], coherent and non-coherent detection methods with massive BS arrays [13], [14], precoding based on instantaneous and long-term statistics [15].

Other papers have also investigated the use of various linear precoders, such as maximum ratio transmission (MRT) and zero-forcing, in massive MIMO URLLC for multiplexing simultaneous devices [16], albeit without accounting for the channel estimation inaccuracy. The authors of [12] have proposed a forward error correction code diversity solution in order to separate pilot-interfering users, by assigning unique user signatures.

Approaches based on stochastic network calculus were adopted in [17] in order to evaluate the latency-reliability trade-off using delay violation probability. The analysis shows that increasing the number of antennas is beneficial in reducing the delay violation probability. The authors of [18] follow a similar approach, where they model the latency-reliability trade-off with a probabilistic constraint on the queue length at the BS. Joint optimization of power, bandwidth, and the number of active antennas for a given number of active devices has been studied [19] in order to maximize the energy efficiency of a massive MIMO network, subject to Quality-of-Service (QoS) constraints.

Power control in a single-cell massive MIMO system has been considered in [20], and the spectral efficiency (SE) has been used as the main metric to develop algorithms based on the weighted minimum SE among the users and the weighted sum SE. Another approach has been taken in [21] where the main metric for optimization is the energy efficiency in the downlink (DL), with QoS constraints. A more recent work [22] extended the max-min SE and max-product SE strategies in [2] by proposing the use of a deep learning approach to predict the optimal power allocation policies from the UE positions.

This paper treats a DL URLLC system, where multiple devices are spatially multiplexed simultaneously. We expand previous knowledge on the trade-offs in terms of SE with insights on what is the corresponding reliability, and the interplay between the two metrics. The work investigates power allocation strategies and their impact on the SE and reliability, for the case of imperfect channel estimation. The number of orthogonal pilots is also varied, in order to quantify the benefit of having a more accurate estimation versus the loss of DoFs for data transmission.
II. SYSTEM MODEL

We consider the DL of a single-cell massive MIMO system where the BS is equipped with $M$ antennas and serves $K$ single-antenna devices; see Fig. 1. The $K$ devices need to satisfy the same latency constraint, which is assumed to be lower than the channel coherence time.

A. Transmission protocol and channel estimation

We assume that a time-division-duplex (TDD) protocol is used with a pilot phase for channel estimation, followed by a data transmission phase. We consider the standard block fading TDD protocol [2, Ch. 2] in which each coherence block consists of $\tau$ channel uses, where $\tau_p$ are used for uplink pilots and $\tau_d = \tau - \tau_p$ for downlink data. We assume that $\tau_p = f K$ where the integer $f$ is the number of pilots per device. We denote $h_k$ the channel vector from the BS to UE $k$ and assume that it is modelled as uncorrelated Rayleigh fading, i.e. $h_k \sim \mathcal{C}\mathcal{N}(0, \beta_k I_M)$, where $\beta_k$ is the large-scale fading coefficient accounting for pathloss and shadowing, defined as [2, Ch. 2]

$$\beta_k = \Upsilon - 10\alpha \log_{10} \left( \frac{d_k}{1\text{km}} \right) + F_k$$

where $\Upsilon$ is the median channel gain at a reference distance of 1 km from the BS. Both $\Upsilon$ and $\alpha$ are parameters computed from established models [2], [23]. $F_k \sim \mathcal{N}(0, \sigma_d^2)$ is a random term modeling the shadow fading as a log-normal random variable.

The uplink pilot sequence of device $k$ is denoted by $\phi_k \in \mathbb{C}^{\tau_p}$ and satisfies $\|\phi_k\|^2 = \tau_p$. The elements of $\phi_k$ are scaled by the pilot power $\sqrt{P}$ and transmitted over $\tau_p$ channel uses. The MMSE estimate of $h_k$ is [2, Ch. 3]

$$\hat{h}_k = \frac{\beta_k}{\beta_k^*} + \frac{\sigma_d^2}{\tau_p P} \left( \frac{1}{\tau_p} \right) \mathbb{N} \phi_k^*$$

where $\mathbb{N}$ is a $M \times \tau_p$ noise with i.i.d. elements distributed as $\mathcal{N}(0, \sigma^2)$ and $\sigma^2 = -174 + 10 \log_{10} (B) + \text{NF}$ [dBm], with NF being the noise factor of the BS. The estimation error $\hat{h}_k = h_k - \hat{h}_k$ has correlation matrix $C_k = \beta_k \left( 1 - \beta_k (\beta_k + \frac{1}{\tau_p P} \sigma_d^2)^{-1} \right) I_M$ [2].

B. Spectral efficiency and precoding design

The BS transmits the DL signal $x = \sum_{i=1}^K w_i c_i$, where $c_i \sim \mathcal{N}(0, \rho_i)$ is the DL data signal intended for device $i$, assigned to a precoding vector $w_i \in \mathbb{C}^M$ that determines the spatial directivity of the transmission and satisfies $\|w_i\|^2 = 1$ so that $\rho_i$ represents the transmit power. An achievable DL for device $k$ can be computed in Massive MIMO by using the hardening bound [1]. This yields

$$\text{SE}_k = \frac{\tau - \tau_p}{\tau} \log_2(1 + \gamma_k) \quad \text{[bit/s/Hz]}$$

(3) with

$$\gamma_k = \frac{\rho_k \mathbb{E} \left[ \|w_k^H h_k\|^2 \right]^2}{\sum_{i=1}^K \rho_i \mathbb{E} \left[ \|w_i^H h_k\|^2 \right] - \rho_k \mathbb{E} \left[ \|w_k^H h_k\| \right] + \sigma^2}$$

(4)

where the pre-log factor accounts for the fraction of samples per coherence block used for DL data. The expectations are computed with respect to the channel realizations. This is not a typical assumption for low-latency transmissions, as the latency constraint can only accommodate one coherence block in time. However, URLLC is expected to use multiple DoFs in frequency, which enables having multiple channel realizations within the available resources. It should also be noted that as the array becomes larger, the channel hardening effect occurs for fewer number of channel realizations taken in the expectation.

The achievable SE in (3) holds true for any precoding scheme. In this work, we select $w_k$ as $w_k = v_k / \|v_k\|$ with $v_k$ being designed according to MR and MMSE precoding [2, Ch. 4]

$$v_k = \left( \sum_{i=1}^K \hat{h}_i h_i^H + \sum_{i=1}^K C_i + \frac{\sigma_d^2}{\tau_p P} I_M \right)^{-1} \hat{h}_k$$

with MMSE.

This choice is motivated by the fact that MMSE is optimal but has high computational complexity. On the other hand, MR is suboptimal but has the lowest complexity among the receive combining schemes.

III. PROBLEM FORMULATION

Several trade-offs arise in the above system model when URLLC requirements are present. The first is about how many channel uses $\tau_p$ should be utilized as pilots for channel estimation for a given coherence block of $\tau$ channel uses. The more are used for pilots, the fewer are left for data transmission. This inevitably increases the communication latency or decreases the reliability of the data transmission phase. Another trade-off comes from the fact that employing power control strategies for increased fairness will lead to lower sum SE in the network. Ideally, in URLLC one would desire to first fulfill the outage requirements, then to achieve the best possible sum SE of the network.

The goal of this paper is to investigate how to control the trade-off between SE and reliability by employing power
control and dimensioning the pilot size, for the MR and MMSE precoders.

A. Outage Definition

We assume that in each latency-coherence block the data transmission rate per device is fixed to $\frac{b}{\tau - \tau_p}$ measured in bit per channel use. By multiplying it with the coherence bandwidth $B_c$, we obtain the threshold rate $R_T = B_c \frac{b}{\tau - \tau_p}$ [bit/s]. Using (3), the maximum achievable rate for device $k$ over the communication bandwidth $B$ is

$$R_k = B \frac{\tau - \tau_p}{\tau} \log_2(1 + \gamma_k) \quad \text{[bit/s]}.$$  

This maximum achievable rate is computed using the hardened DL SINR $\gamma_k$ over the multiple DoFs in frequency in (4). The outage probability of device $k$ measures the probability that the DL SINR $\gamma_k$ cannot support the transmission of the fixed packet of $b$ bits within the latency constraint. It is defined as

$$P_{\text{out}}^{\text{device}} = \Pr[R_k < R_T] = \Pr \left[ B \frac{\tau - \tau_p}{\tau} \log_2(1 + \gamma_k) < B_c \frac{b}{\tau - \tau_p} \right]$$

and it is evaluated across the device positions. In addition to the device outage, we can define the system outage, which is the probability that at least one device of a given setup would be in outage. This corresponds to measuring the worst user outage across the device deployments. Its expression is given by

$$P_{\text{out}}^{\text{sys}} = \Pr \left[ B \frac{\tau - \tau_p}{\tau} \log_2(1 + \min_k \gamma_k) < B_c \frac{b}{\tau - \tau_p} \right].$$

B. Power Allocation

The DL transmit powers $\{\rho_k : \forall k\}$ in (3) need to be selected. To achieve high reliability, it is important to increase the fairness between the devices in the network, meaning that devices located further away from the BS, which experience worse channel conditions, should be allocated more power. This induces a trade-off, between how much power should be allocated to devices experiencing a weak channel, at the expense of a performance degradation of the network sum SE and of devices with stronger channels. To this extent, several power allocation strategies have been proposed in the literature [2], [22], [20], [21]. The most simple solution is equal power allocation, i.e., $\rho_i = P_{\text{max}}/K$ with $P_{\text{max}}$ being the maximum DL transmit power. Alternatively, two prominent examples are the max-min fairness and max-product SINR strategies, which can be mathematically formalized as follows:

$$\max_{\{\rho_k : \forall k\}} \min_k \text{SE}_k$$

subject to $\sum_{k=1}^{K} \rho_k \leq P_{\text{max}}$

and the max-product SINR, given by

$$\max_{\{\rho_k : \forall k\}} \prod_{k=1}^{K} \gamma_k$$

subject to $\sum_{k=1}^{K} \rho_k \leq P_{\text{max}}$.

The max-min fairness in (9) provides complete fairness by only counting the SE achieved by the weakest device in the network. This results in the same SE for all; that is, a device has no benefit of having a good channel condition. This inevitably reduces the sum SE. The max-prod power allocation policy (10) balances between sum SE and fairness. The impact of these schemes has been mostly studied from the perspective of average network SE. We extend this result to the previously defined metrics of device outage and system outage, and show the interplay between achieving reliability and SE.

IV. Numerical Analysis

The spectral efficiency, device outage and system outage are evaluated using Monte-Carlo simulations. Device deployments are generated based on the scenario defined in Section II. The simulation parameters are described in Table I. The number of symbols in a coherence block is based on the size of the coherence bandwidth, which here is assumed to be 100 kHz. The number of channel instances used in taking the expectation in (4) is given by the number of coherence blocks in the total bandwidth, and is equal to $B/B_c$.

The network sum SE of the three power allocation strategies for the MRT and MMSE precoders are shown in Fig. 2. As expected, the sum SE for the max-min SINR is the lowest for each the MRT and MMSE precoders, as it introduces a large amount of fairness in the system, penalizing the SE of strong devices for the benefit of the weak ones.

It is interesting to observe that the max-product SINR method performs nearly identical to the equal power allocation. This result might seem somewhat surprising, because one would expect the max-product SINR to provide some degrees of fairness. However, when the CSI is imperfect, the weak devices exhibit worse CSI accuracy, and by allocating more power to inaccurate precoders, the system would create more
interference, thereby reducing SINR of other devices and the product of SINRs. Therefore, max-product SINR, paired with imperfect CSI, performs similarly to equal power.

The outage is shown in Fig. 3. Firstly, it can be again noticed that the max-product SINR strategy performs nearly the same as equal power allocation, achieving almost identically device and system outage, at an increased complexity of computation. Secondly, the improvements in fairness of the max-min SINR strategy result in a tremendous decrease of the system outage, which is crucial for URLLC. Moreover, the device outage is also decreased by a small margin, compared to equal power and max-product SINR.

Based on the SE and outage results from Fig. 2-3, it can be noticed that the precoder choice of MMSE over MRT has a significant improvement on the sum SE, while bringing only a minor detriment to the outage. Moreover, the results show that the sum SE needs to be sacrificed in order to obtain the considerable improvement in outage provided by the max-min SINR strategy, if the system is to operate at its peak reliability. Fig. 4 shows the decrease in sum SE when employing two pilots per device for CSI, as the number of devices grows. Fig. 5 shows that from the outage perspective, investing an extra pilot per device for CSI provides further improvements.

In addition, we show the shape of the probability density functions of the device SINRs ($\gamma_k$) in Fig. 6. One first observation is that the distribution of the max-min SINRs has a smaller lower tail compared to the equal power, which is why the outage is also lower. Secondly, it can be seen that the max-min SINR strategy does not achieve the same high SINRs as equal power, due to the fairness. This is the reason why the sum SE of max-min experiences a dramatic decrease (Fig. 4, note that the X-axis of the SINR is in dB).
Fig. 6. PDFs of the SINRs for MMSE, with equal power and max-min SINR allocation.

V. CONCLUSION

This paper investigated the means of achieving ultra-reliable low-latency communications in a single-cell downlink massive MIMO system. More specifically, we utilized state-of-the-art power allocation strategies and precoders, in order to show novel results in terms of outage. Moreover, we exposed two trade-offs: between achieving high SE and high reliability by power allocation strategies; and between allocating a single uplink pilot symbol versus two pilot symbols per device, for the case of multiplexing several devices in a low-latency setting.

We concluded that the max-min SINR strategy is the best in terms of outage, due to its increased fairness among devices. However, it leads to a much lower sum SE due to allocating more power to the weaker devices. The max-product SINR strategy achieves nearly identical performance as the equal power, at an increased complexity of computation, making it the less suitable scheme in this scenario with imperfect CSI.

Furthermore, the results showed that the precoder choice is not highly important for the reliability, but choosing MMSE can considerably improve the sum SE with a minimal decrease in reliability.

The use of an extra pilot symbol per device proved to be beneficial in terms of outage for this particular low-latency setting, despite the decrease in terms of sum SE for the case of larger number of devices.

REFERENCES


ACKNOWLEDGMENT

The researchers from Aalborg University have been supported partly by the European Research Council (ERC Consolidator Grant nr. 648382 WILLOW) and by the Danish Council for Independent Research (DFF) (CELEST project). L. Sanguinetti was supported by Pisa University under the PRA 2018-2019 Project CONCEPT.