Key Solutions for a Massive MIMO FDD System

Wolfgang Zirwas
Nokia Bell Labs
Munich, Germany
Email: wolfgang.zirwas@nokia-bell-labs.com

Mikael Sternad, Rikke Apelfröjd
Signals and Systems
University of Uppsala, Uppsala, Sweden
Email: Mikael.Sternad@signal.uu.se

***Abstract*—The ongoing standardization within 3GPP for the so called new radio (NR) system has identified massive multiple-input multiple output (MIMO) transmission, also called full dimension MIMO, as one of the main contributors to higher spectral efficiency for the mobile broadband case. In particular for radio frequencies below 6 GHz, channel estimation has to be supported in frequency division duplex (FDD) as well as time division duplex (TDD) operation. In TDD we may obtain downlink channels by estimating uplink channels, assuming reciprocity. For FDD, codebook based design as well as some type of explicit feedback is under discussion. Separately, there are also ongoing discussions of the question if massive MIMO in combination with FDD is a reasonable choice at all.**

Here we highlight some of our recent results obtained within several 5G research projects. To our understanding they overcome some of the inherent limitations of massive MIMO for FDD. As indicated by simulations, the resulting concept enables a grid of beam (GoB) and reference signal design with a reasonable downlink reference signal overhead of around 10 percent, together with reasonable feedback overhead of several hundred kbit/s per UE. Such a design attains around 90 percent of the massive MIMO system performance with ideal channel state information.

Keywords — massive MIMO; channel estimation; CoMP; pilot contamination; beam management

I. Introduction

There is an ongoing debate about the possible and preferable ways to exploit the benefits of massive multiple-input-mupltiple output (MIMO) antenna arrays. The original downlink concept is based on the usage of time division duplex (TDD) and the assumption of channel reciprocity, while a frequency division duplex (FDD), grid of beam concept has been proposed in, e.g., [1].

Recently, an interesting and thought-provoking input to this debate has been provided by a paper presently still under review [2]. It evaluates the use of various FDD-based massive MIMO schemes by using measured channels from a test-bed in Lund, Sweden. It concludes that FDD massive MIMO grid of beam (GoB) systems are not robust with respect to the investigated scenarios and often perform far from the theoretical performance bound. This paper, in line with several other research groups, see e.g. [3], regard the use of TDD as the only feasible solution.

However, on closer inspection, the massive MIMO setting creates not only challenges, but also properties of the channels that can be used to overcome these challenges. In particular, a typical radio channel can be made sparse in the spatial domain. This allows to consider new solutions that overcome some of the main concerns regarding the use of massive MIMO in FDD. It also enables the cooperation of a still larger number of antenna elements at multiple sites, a combination of massive MIMO and coherent coordinated multipoint transmission (JT-CoMP), or network MIMO [1].

We begin this paper in Section II by briefly recapitulating the main concerns with respect to FDD, as expressed for example in [2], and complement with similar observations from our measurements and system level simulations. In particular, non line of sight (NLOS) urban macro scenarios are characterized by multiple reflections. Therefore, under the assumption of a grid of beam (GoB) design, downlink users (UEs) receive not only one strong beam, but instead many so called relevant channel components that are above a power threshold. These relevant channel components vary in case of MU MIMO for different UE locations and would have to be estimated as well as reported individually for all scheduled UEs, with sufficiently high quality. This creates both a reference signal overhead problem since many channels have to be estimated simultaneously, and also an uplink feedback overhead problem.

We then in Section III present tools and ideas for overcoming the most severe issues, with a main focus on the required overhead due to downlink reference signals for obtaining the channel state information (CSI RS). We introduce our concept of $coded\ CSI\ RS$, which acts as a game changer: Using this design, the reference signal overhead just scales with the typical number of relevant channel components to be estimated per UE, instead of with the total number of antenna elements or beams. In particular, the overhead does not scale approximately linearly with the number of simultaneously served UEs as is the case in, e.g., [3]. A reference signal overhead that scales with the number of UEs would also be a problem for several recently suggested schemes for the reduction of the downlink reference signal overhead in FDD massive MIMO, see e.g. [4], [5], [6] and [7].

A massive network MIMO GoB framework, developed recently within the EU H2020 Fantastic5G project, [1], will be presented and utilized. We will extend this framework in Section IV by providing a novel beam deactivation scheme. It simultaneously improves the CSI estimation quality and reduces the CSI reporting overhead. A simulation that includes appropriate user grouping, scheduling, CSI estimation, reporting of relevant channel components and taps as well as the beam deactivation, is then presented in Section V. It indicates a high potential for massive MIMO FDD systems. Due to page limits, we refer to suitable references for many details of the involved components of the solution.

II. Challenges for a massive MIMO FDD System

There are good reasons that the original concepts for massive MIMO as well as for JT CoMP have been developed for TDD systems. Assuming a high number of downlink transmit antennas, either at a single or at a distributed antenna array, typically requires a high number of orthogonal CSI RSs in FDD. It would also require a high uplink overhead for reporting the estimated CSI from the UEs to the base station (denoted next generation NodeB, or gNB, in evolving 5G standards). In contrast, the use of ideally just a single uplink sounding reference signal (SRS) per UE and per radio resource block would allow the estimation of all uplink radio channels to all antenna elements. In TDD systems, these might then be reused for the downlink due to the reciprocity of downlink and uplink radio channels.

In FDD, channel reciprocity cannot be assumed. Downlink channel estimation by the use of orthogonal downlink CSI RSs, one per antenna element, here encounters a fundamental limitation: The channel coherence time and the coherence bandwidth will set an upper bound for the number of orthogonal CSI RSs. In addition, there will be an optimum balance of CSI RS overhead versus user data transmission within such a (block fading) transmission block. This tradeoff is between reporting of more relevant channel components with higher accuracy versus reducing the number of resource elements available for data transmission, thus limiting the maximum data throughput.

A first countermeasure to this problem is to use a transformation from the sets of antennas into a beam space at the gNB, the so called GoB concept. The number of beams may be equal to, larger or smaller than the number of antennas. If we chose to use equal or fewer beams than number of antennas, then the number of channel components that are strong for a user, and therefore need to be estimated, is reduced, due to the directed transmission of beams.

The balance here is that the use of too few beams will reduce the attainable performance and the multiuser MIMO scheduling gain. We will in the following assume that the number of beams is so high that each user typically receives strong signals from several beams, denoted relevant beams, or relevant channel components.

High cell capacity requires spatial multiplexing of multiple streams to multiple UEs. Each UE $k \in \mathcal{N}^K$ will receive $N_{rel,k}$ beams above a specified power threshold. The number N_{RS} of required orthogonal CSI RSs will then scale with $\max(N_{rel,k})$ and with K. In the from an overhead perspective best - but unlikely - case, all UEs would have the same N_{rel} beams as relevant channel components.[1] In the opposite extreme case, each UE would have a different set of relevant beams so that $N_{RS} = \sum_k N_{rel,k}$ orthogonal reference signals would be required. A typical situation will be somewhere in between. Even with a GoB, the use of a minimal number N_{RS} of orthogonal CSI RSs would therefore still require a large overhead. It would also require adaptation of the transmitted CSI RSs to the set of currently scheduled UEs.

The results in [2] are based on measured radio channels. Block wise fading and allocation of the best fitting set of orthogonal CSI RS for various FDD scenarios is assumed. Due to the CSI RS overhead, in some cases only 40% to 50% of the capacity with perfect CSI is then attained. This, obviously, would be a strong drawback for a FDD system.

Generally, these observations are in line with our own evaluations. For example in the study [8] of the GoB concept, $N_{rel,k}$ varied from few to several tens of relevant channel components per UE. In addition the relevant channel components were different from UE to UE, depending on the user location.

A related problem is the CSI reporting overhead. For 3GPP LTE and new radio full dimension MIMO, the focus is on codebook based feedback schemes for FDD. For a future phase II of the 3GPP New Radio standardization, also explicit CSI feedback is being considered, as it would result in improved precoder performance with reduced inter-stream interference. However, the related CSI feedback overhead has to be considered carefully. For example, reporting explicit CSI for 40 channel components every 5 ms with a quantization of 10 bit for each subband of 6 physical resource blocks (PRB) of a 20 MHz = 100 PRB system would result in a feedback rate greater than 1 Mbit/s for each UE.

An additional challenge has been identified in [2]: For clusters of closely spaced UEs, or a hot spot scenario, the performance with the GoB concept is reduced. This problem is caused by using a fixed allocation of a low to moderate number of beams that are designed to span the full cell area. The number of beams serving a hot spot area might then become small and inadequate.

[1] However, as noted in the discussion on power normalization loss below, such a case would be the worst with respect to the possibility to design a linear joint precoder with good properties.

To conclude, although a GoB system can alleviate the downlink reference signal overhead problem in FDD by using a smaller number of beams than the number of transmit antennas, there remain several challenges: The downlink reference signal overhead is still large with orthogonal reference signals, the total feedback load with many users will be large, and the number of beams covering a given area may be inadequate, when using a fixed positioning of beams.

III. How to design a massive MIMO FDD system

The introduction of an advanced interference mitigation framework can become a main differentiator for New Radio as compared to LTE. This motivates us to investigate our solution in a setting where JT CoMP is used over a cooperation area spanning several cells. For example, our evaluation case, described in Section V below, consists of nine adjacent cells located at three sites, each equipped with a massive MIMO antenna array, with in total N_{beam} beams per cooperation area, where $N_{beam} = 288$ will be used for evaluation.

Despite the high number of beams, our aim will be to achieve close to 90 percent of the ideal perfect-CSI massive network MIMO performance. The reference signal overhead should be about 10 percent, together with reasonable feedback overhead of about one to several 100 bit/subframe of 1 ms or, equivalently, a few 100 kbit/s per UE. See for example the final deliverable from the EU project Fantastic5G [1].

a) Use of a Grid-of-Beams Downlink

A first step towards an efficient FDD solution is to exploit the spatial sparsity of a well designed GoB system. The number of relevant channel components will then be limited for each particular user, with $N_{rel,k} \approx 10$ to several tens of relevant channel components per UE, even in the case of a cooperation area with, e.g., $N_{beam} = 288$ potential beams. This is the result of the directed transmission of the beams, as compared to omni-directional single antenna elements.

To avoid problems with hotspots of users, the beam pattern should be adaptable to the traffic demand on a slow time scale. Spatial clusters of UEs can then be served by more beams than scheduled UEs.

b) Use of non-orthogonal coded downlink CSI RS

The next crucial step is to make the best use of the limited number of CSI RSs within the limited coherence time-frequency region. Here, the use of so called coded CSI RSs [9], where non-orthogonal CSI RSs are allocated to antenna ports (AP) or beams, can become a game changer.

Briefly, instead of allocating one orthogonal CSI RS for each of the N_{beam} beams, so that $N_{RS} = N_{beam}$, each beam instead transmits a non orthogonal CSI RS sequence of much shorter length. These non orthogonal sequences of length $N_{RS}^{coded} \ll N_{beam}$ are unique per beam. They should have the property to allow CSI reconstruction at the UEs of almost all arbitrarily selected sets of up to N_{RS}^{coded} relevant channel components. Each UE will then be able to estimate its individual set of up to N_{RS}^{coded} relevant channel components. For the overall CSI RS design, this means that the required overhead now scales with the numbers $N_{rel,k}$ of relevant channel components. Details and evaluations of the concept of coded CSI RS can be found in [9] and in [10].

For example, assuming $\max(N_{rel,k}) = 40$ relevant channel components per UE, a sequence length of 40 reference symbols is sufficient to allow all UEs to uniquely determine their UE-individual sets of relevant channel components. Ray tracing models and system level simulations have indicated that the number $\max(N_{rel,k}) = 40$ is a reasonable assumption [9]. The required $N_{RS}^{coded} = 40$ would then equal the number of resource elements that is today used per PRB for CSI RSs in LTE systems. The corresponding overhead would then be only 4.7 percent with one CSI RS transmission every 5 ms. This overhead might increase to around 10 percent in case higher estimation accuracy or a higher CSI RS transmission rate is targeted.

Features that increase the spatial sparsity would help to further decrease the overhead and/or to improve the estimation performance. For example, if UEs are equipped with eight receive antennas, then a suitable receive beamformer can on average reduce the number of relevant channel components by about 50% [8].

The important aspect here is that for spatially sparse channels generated by massive MIMO GoB, we do not need orthogonal CSI RSs: Suitably designed non-orthogonal sequences of limited length are sufficient.

c) Use explicit CSI feedback for relevant channels only, and use feedback compression

To overcome the feedback challenge, we propose the use of explicit CSI feedback per *relevant* channel component, with compression that takes the statistics of the time/frequency selective channel into account.

Reporting is then limited to channel components above a pre-set power threshold. For a properly defined reporting threshold in the range of, e.g., 20 to 25 dB below the strongest channel component, the unreported channel components can be set to zero at the gNB when forming the downlink beamformers. The inter-stream interference due to the precoder mismatch will remain small.

As outlined in [11], we may let each UE report semi-statically, e.g., every 500 ms, the set of its relevant channel components as well as the significant taps of the channel impulse response. A relatively high overhead will then be generated relatively rarely for the identification of the relevant channel components and taps. Then, normalized estimates of the significant taps only are reported, with a rate adjusted to the fading and the transmission requirements.

If the channel is sparse in the time domain, then the reporting of only the significant time-domain taps of the channel, as suggested in [11], will be more advantageous than source-coding the frequency-selective frequency-domain channels. In [12], the explicit feedback is evaluated in a similar manner using compressed sensing, demonstrating its potential for sparse channel impulse responses.

Again, the key to success is here to exploit sparsity, both spatially and temporally, and to design the overall system to have a sparse channel structure.

Analysis for measured urban macro radio channels indicate that typically there are on average around 20 relevant time domain taps per channel component. Even without bit loading and assuming a ten bit quantization per tap transmitted every 5 ms, around 40 kbit/s would then be required per channel component. For 10 to 40 relevant channel components, a feedback overhead in the range of several hundred kbit/s per UE seems attainable. One should note that this case covers frequency selective CSI information for the full bandwidth of, e.g., 20 MHz. Lower subband sizes will generate lower overhead per UE.

The high-level downlink transmission concept

Figure 1 provides a high level illustration of the proposed concept, comprising the following components

1) The use of massive MIMO.
2) Spatial multiplexing to more than 10 UE/cell.
3) Use of network MIMO or inter site JT CoMP precoder $\mathbf{W}$ over the relevant channel components generated by the GoB beamformer matrices $\mathbf{V}$.
4) The use of coded CSI RS for accurate channel estimation with limited overhead for CSI RSs.
5) Beamformers on the UE with up to 8 antenna elements that increase the sparsity of the overall downlink channel matrix $\mathbf{H}$, which includes the transmit GoB and receive filters.
6) CSI reporting per relevant taps and relevant channel components minimizing the feedback overhead to the relevant information. Channels are predicted for moving UEs, see [10], [13].
7) Use of interference floor shaping to reduce inter-cluster interference, e.g., based on the cover shift concept as well as the tortoise concept [14].

As indicated in Figure 1, ideally the target performance is a factor 7 - 10 higher spectral efficiency as compared to a LTE 4x4 system, while keeping a reasonable overhead for CSI RSs of about 5% - 10% as well as a reasonable overhead for CSI feedback.

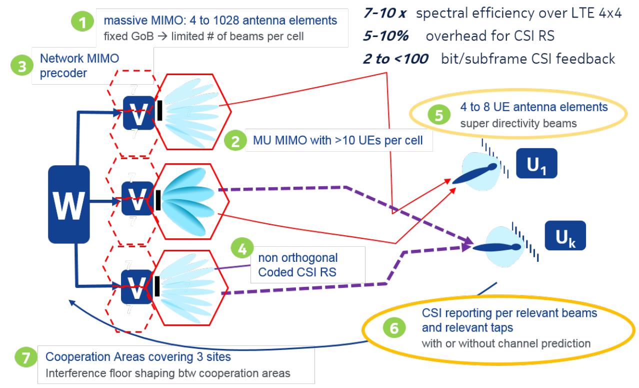

Fig. 1: Interference Mitigation framework based on massive MIMO and JT CoMP in an FDD setting.

IV. Beam Deactivation for Massive MIMO FDD Systems

Sparsity is the key for attaining a workable FDD massive MIMO system. So why then send CSI RS in beams that are not relevant to any UE? Dynamic beam deactivation would lower the CSI RS power. This improves the channel estimation accuracy when used in combination with coded CSI RSs, since it reduces the inter-code interference due to the code non-orthogonality. It will also improve the power efficiency and reduce the inter-cell interference and also the CSI feedback overhead. We will here discuss an optimum beam management or beam deactivation scheme as an extension to the overall massive MIMO GoB framework of Figure 1.

For a single UE, beam deactivation would be quite straightforward, for example by switching off all irrelevant beams, i.e., those of the transmit beams that would be received below a specified power threshold, on average over the short-term fading.

With many active UEs per cell, the situation becomes more complicated as an irrelevant beam for one UE might then be a serving beam for another UE. Analysis for typical randomly scheduled UEs in the system simulation setting of Section V, with 288 beams and 80 users, indicates a strong interconnection between beams and UEs. For example, in the vertex diagram of Figure 2, each UE is connected to Tx-beams from different cells and sites in a quite irregular manner. The set of active beams therefore needs to be adapted to the set of scheduled UEs. Due to frequency dependent multi-user scheduling, such a beam deactivation should be subband specific.

Figure 3, which shows an example of sorted received beam powers, provides some further insights:

- First, the number of relevant channel components - or beams - vary significantly from UE to UE. While UEs at the cell center are mainly served by a few relevant beams, UEs at the border to several cells might receive 50 or more beams as relevant channel components. For that reason, a simple CSI feedback framework with a fixed number of reported channel components seems to be inadequate.

- Second, almost all UEs experience an interference floor where a high number of beams are received with a power close to the relevant channel component power threshold P_{TH}. These beams will not be reported, but due to their high number of 50 or more channel components, the sum of their (neglected) powers would generate a strong inter stream interference for a fully loaded system.

- Third, there are Tx-beams, which are received well below the power of the relevant channel components - like minus 30 to 40 dB below the strongest channel component - and therefore do not affect the performance of a zero-forcing (ZF) precoder at all.

In Figure 4 these three regions are illustrated with the relevant, the interference floor and the irrelevant channel beams.

A further challenge is that some of the UEs have the same beam, or subset of beams, as their strongest serving beam(s). This can be concluded from a deeper analysis of the vertex diagram in Figure 2. The typical high correlation of these radio channels will often lead to a high power normalization loss (PNL) for zero-forcing-like linear precoders and this PNL will degrade the overall performance.

To alleviate this problem, the UEs that would generate a very high PNL if admitted - even with all beams active - should be removed from the served multi-user set. A separate subband with a different setup of active beams should be used to serve these users.

For the remaining UEs, the best zero forcing precoding performance (under the assumption of ideal CSI) is attained when all N_{beam} beams are active, as this will lead to the lowest condition number of the overall channel matrix $\mathbf{H} \in \mathcal{C}^{KxN_{beam}}$. However, the corresponding overhead for CSI RSs and CSI reporting for such a large cooperation area can be reduced. Together with the observations from Figure 4, we end up in a multidimensional non-convex optimization problem, with the following, partly conflicting, goals:

We should find the best beam deactivation pattern for a defined set of scheduled UEs, such that

- the number of beams that contribute to the interference floors of the different UEs should be reduced,

- also, the number of relevant channel components should be reduced to lower the feedback overhead, but

- this has to be done while causing a not too large extra power normalization loss, as compared to the case when all beams are active.

It is proposed to use an iterative optimization of a weighted cost function:

$$\begin{aligned}
\{\mathcal{B}_{opt}|\mathcal{K}_{sched}\} &= \arg\min(\alpha_1 PNL(\mathcal{B}_l) \\
&\quad +\alpha_2 N_{rel}(\mathcal{B}_l) + \alpha_3 N_{IF}(\mathcal{B}_l)) \ ; \\
PNL(\mathcal{B}_l) &= \sum \|w_{ij}\|^2; \ with \, w_{ij} = \mathbf{W}^l_{ij}; \\
\mathbf{W}^l &= pinv(\mathbf{H}_{beam}(\mathcal{B}_l, \mathcal{K}_{sched})) \ ,
\end{aligned} \tag{1}$$

with α_1, α_2 and α_3 being scalar weight factors, defining the relative importance of the PNL, the number N_{rel} of relevant channel components and the number N_{IF} of interference floor channel components. Here, $\mathcal{K}_{sched}$ is the currently scheduled set of UEs, $\mathbf{H}_{beam}(\cdot,\cdot) \in \mathcal{C}^{KxN_{beam}}$ is the overall channel matrix for all N_{beam} beams and all K UEs, but with rows and column set to zero according to the beam deactivation and scheduling assumed in iteration l. Then, with pseudoinverse $pinv(\cdot)$, $\mathbf{W}^l$ is the joint linear precoding matrix at iteration l.

Here, $\mathcal{B}_l$ is the combination or set of active beams $\{b_1 \cdots b_{N_{beam}}\}$ at iteration l, with $b_l \in 0, 1$, i.e., $b_l = 1/0$ indicating an active or inactive beam.

Also, $N_{rel}(\mathcal{B}_l)$ is the overall number of relevant channel components, for which the Rx power of beam j at some UE i is $P_{Rx}(\mathbf{H}_{ij}) > P_{TH}$.

Finally, N_{IF} is the number of channel components forming the interference floor, for which $P_{TH} \geq P_{Rx}(\mathbf{H}_{ij}) > P_{TH,IF}$ for some UE i, where $P_{TH,IF}$ is the threshold power for the interference floor. Note, $\mathbf{H}_{ij}$ - and accordingly $\mathbf{H}_{beam}$ - is column wise normalized to the strongest beam per UE .

In the iterative approach, beams are deactivated one by one. At each iteration l, the beam that generates the largest decrease of the cost function when removed will be deactivated. The final optimum beam set $\mathcal{B}_{opt}$ is then obtained, for example, by defining a maximum allowed PNL value PNL_{max} and choosing the beam set that generates a PNL just below PNL_{max}. The reduction of the number of relevant channel components brings a benefit due to the reduced feedback overhead. The reduced number of interference floor channel components furthermore significantly reduces the inter-stream interference. Depending on the situation, this often outweighths the additional PNL loss as compared to full beam activation due to the resulting reduced inter-stream interference.

V. Simulation results

The proposed concept has been simulated for a nine cell cooperation area with a massive MIMO antenna per

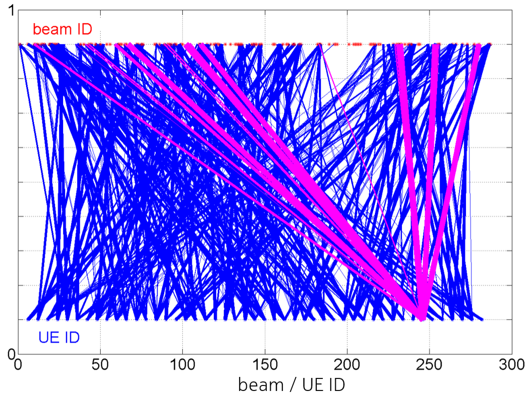

Fig. 2: Vertex diagram representing the downlink signal powers from 288 gNB beams (red stars at top line) to 80 UEs (at the bottom line); the thickness of lines indicate the relative average power of the link. Channels to the UE no. 60 are highlighted by magenta lines.

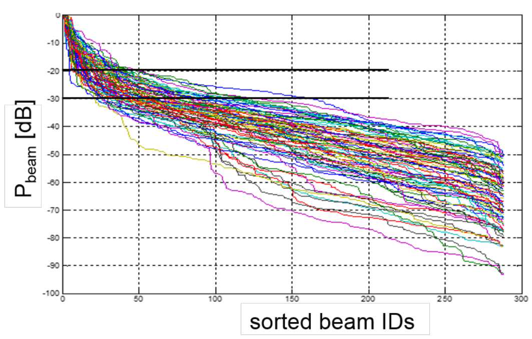

Fig. 3: Rx- beam power in [dB], relative to the strongest received beam, sorted by decreasing received power. Simulation results per UE for 80 UEs in a nine cell cooperation area. The upper horizontal black line indicates the relevant channel component threshold P_{TH}, here at -20 dB, while the lower line indicates the limit $P_{TH,IF}$, here at -30 dB.

cell of 32 x 16 elements per polarization, i.e., for overall 1024 antenna elements, forming eight horizontal and two vertical fixed beams per polarization, i.e., overall 32 beams per cell. Inter-site cooperation is done for three adjacent sites or nine cells, leading to the already discussed 288 beams per cooperation area.

The basic simulation parameters are close to the 3GPP case I simulations with an RF-frequency of 2.6 GHz, an inter-site distance of 500 m, an outdoor-to-indoor penetration loss of typically 20 dB, an UE noise figure of 7 dB, and a Tx-power per cell of 49 dBm for a 20 MHz bandwidth. In addition, the UEs are equipped with an eight element uniform linear array and apply a maximum ratio combiner to their strongest Rx-beam. A regularized zero forcer per cooperation area is designed to ensure a low inter-stream interference within the cooperation area with ideal CSI.

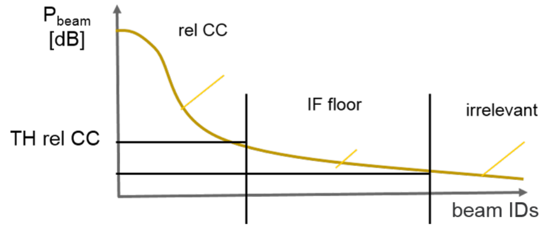

Fig. 4: Typical distribution of relevant and interfering channel components, where "Th rel CC" represents P_{TH}, the threshold for relevant channel components.

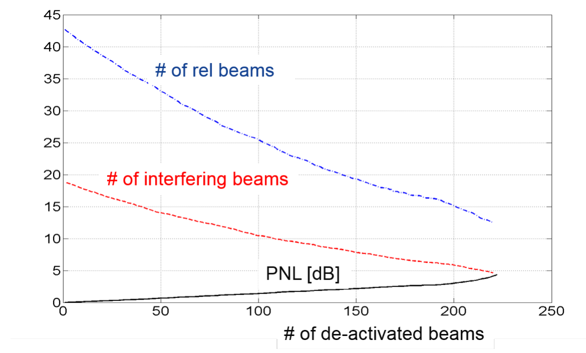

Fig. 5: Average number of relevant (blue, dash dotted) and interfering (red, dash) channel components and the resulting power normalization loss in [dB] (black, solid) as a function of the number of deactivated beams.

We first investigate the iterative progress of the beam deactivation described in Section IV. Starting with 80 users randomly placed within the cooperation area, the result of successively deactivating one beam per iteration is seen in Figure 5 for up to 220 iterations. At the last iteration, only $288-220 = 68$ out of 288 beams remain activated. In this case the scheduler has removed 16 out of the 80 UEs, leaving 64 active UEs. A clear reduction of relevant and interfering channels is here produced, at the cost of a PNL increase of up to 5 dB.

We now investigate the performance of the overall concept as outlined by Figure 1. The use of coded CSI reference signals for channel estimation is combined with reporting of only the estimated relevant channel component and taps, together with beam deactivation, leading to a joint regularized zero forcing precoder. The resulting performance is compared to a case with ideal CSI, that uses all beams.

In one exemplary simulation, we had 172 randomly placed UEs within the cooperation area. Of these, 63 were removed by the scheduler, as they would cause a large PNL.[2] For the remaining 109 UEs, the beam deactivation according to Section IV has been applied.

[2]The main reason for the high PNL is that several UEs would have the same beam as strongest serving beam. Such UEs should then instead be served in orthogonal time or frequency resources. This principle is in line with scheduling guidlines and results in [13].

It results in the remaining use of 149 of the 288 beams, i.e., more than 50% deactivation, where the resulting PNL increase has been limited to 3 dB. After beam deactivation, another 19 of the 109 UEs were set aside by the scheduler, leaving 90 UEs, or on average ten UEs per cell, to be scheduled simultaneously.

Coded CSI reference signals were here used with a long sequence length of $N_{RS}^{coded} = 60$ All channels were furthermore estimated based on two sequences, which improved the estimation accuracy [9]. This resulted in a reference signal overhead of 14 percent. The relevant channel components were fed back every 5 ms with adapted quantization levels per relevant tap, resulting in an average feedback rate per UE of 450 kbit/s. The downlink gross spectral efficiency (without overhead) then became 40.6 bit/s/Hz/cell, corresponding to an average rate of 4.1 bit/s/Hz per user.

With ideal CSI, a gross spectral efficiency of about 46 bit/s/Hz/cell could have been attained. The resulting average data rate per UE would then be 4.64 bit/s/Hz. Accordingly, the gross spectral efficiency based on estimated and quantized CSI was degraded by only about 10 percent as compared to the case of ideal CSI.

As reference case, we use the 3GPP LTE 4x4 MIMO performance, which under the present assumptions, ideal CSI and 43% overhead has spectral efficiency of around 3 bit/s/Hz/cell, or 5.2 bit/s/Hz/cell without overhead. The case investigated above improves on this by a factor 9.

Additional simulations, evaluations and fine-tuning of the components of the overall scheme are needed and ongoing. It is likely that more suitable performance-complexity tradeoffs can be found, but the results above clearly indicate the potential of massive MIMO FDD systems with gross spectral efficiencies around 50 bit/s/Hz/cell. This is almost a factor of ten higher than what can be achieved with a 4x4 MIMO LTE system.

VI. CONCLUSION

We have outlined an overall framework for FDD massive MIMO systems that includes the use of grid-of-beams and inter-site JT CoMP for interference mitigation. The resulting gross spectral efficiency attains around 90 percent of the performance for ideal CSI.

About a factor of ten higher spectral efficiencies over LTE 4x4 MIMO seems to be feasible by combining the proposed concepts, at a reasonable CSI RS overhead of about 10 percent in typical urban macro scenarios. The feedback overhead becomes high but not unreasonable. In the simulation example it was in the range of several hundreds of kbit/s per UE at a continuous reporting rate of 200 Hz. This might be reduced, e.g., by using reliable channel prediction.

From this perspective, we believe that massive MIMO in FDD - especially in combination with JT CoMP - has a strong potential and should be further considered for a 3GPP new radio system, as is the case at present. It might furthermore be worth to consider a coded downlink CSI RS plus reporting on relevant channel components and taps scheme also for TDD downlinks, typically as an add on to SRS reciprocity based channel estimation.

ACKNOWLEDGMENT

This work has been performed in the framework of the Horizon 2020 project FANTASTIC-5G (ICT-671660) receiving funds from the European Union. The authors would like to acknowledge the contributions of their colleagues in the project, although the views expressed in this contribution are those of the authors and do not necessarily represent the project.

REFERENCES

[1] B. S. et. al., "Fantastic5G deliverable D4.2," *EU H2020-ICT-2014-2 Fantastic5G, http://fantastic5G.eu*, April 2017.

[2] J. Flordelis, F. Rusek, F. Tufvesson, E. G. Larsson, and O. Edfors, "Massive MIMO performance - TDD versus FDD: What do measurements say?" *ArXiv:1704.00623v1 [cs.IT], 3 April 2017*, 2017.

[3] T. L. Marzetta, "Massive MIMO: An introduction," *Bell Labs Technical Journal*, vol. 20, pp. 11–22, 2015.

[4] J. Choi, D. Love, and P. Bidigare, "Downlink training techniques for FDD massive MIMO system: Open-loop and closed-loop training with memory," *IEEE Journal of Selected Topics in Signal Processing*, vol. 8, pp. 802–814, 2014.

[5] S. Noh, M. Zolowski, Y. Sung, and D. Love, "Pilot beam pattern design for channel estimation in massive MIMO systems," *IEEE Journal of Selected Topics in Signal Processing*, vol. 8, pp. 787–801, 2014.

[6] Y. Han, J. Lee, and D. Love, "Compressed sensing-aided downlink channel training for massive MIMO," *IEEE Trans. on Communications, DOI 10.1109/TCOMM.2017.2691700*, vol. 8, pp. 802–814, 2017.

[7] C. Tseng, J. Wu, and T. Lee, "Enhanced compressive downlink CSI recovery for FDD massive MIMO systems using weighted block ℓ_1-minimization," *IEEE Trans. on Communications*, vol. 64, pp. 1055–1067, 2016.

[8] R. S. Ganesan, W. Zirwas, B. Panzner, K. Pederson, and K. Valkealahti, "Integrating 3D channel model and grid of beams for 5G mMIMO system level simulations," *IEEE Vehicular Technology Conference, VTC-2016-Fall*, September 2016.

[9] W. Zirwas, M. B. Amin, and M. Sternad, "Coded CSI reference signals for 5G exploiting sparsity of FDD massive MIMO radio channels," *20 International ITG Workshop on Smart Antennas*, March 2016.

[10] R. Apelfröjd, W. Zirwas, and M. Sternad, "Joint reference signal design and kalman/wiener aided channel estimation for FDD massive MIMO. extended version," *http://www.signal.uu.se/Publications/pdf/r1701.pdf*, June 2017.

[11] T. Wild, C. Hoek, G. Herzog, and J. Koppenborg, "Multi-antenna OFDM channel feedback compression exploiting sparsity," *European Wireless*, 2013.

[12] G. Wunder, H. Boche, T. Stroehmer, and P. Jung, "Sparse signal processing concepts for efficient 5G system design," *IEEE Access*, June 2015.

[13] R. Apelfrojd and M. Sternad, "Design and measurement-based evaluations of coherent jt comp: a study of precoding, user grouping and resource allocation using predicted csi," *EURASIP Journal on Wireless Communications and Networking*, vol. 2014:100, June 2014.

[14] L. Thiele, M. Kurras, S. Jaekel, and W. Zirwas, "Interference floor shaping for liquid coverage zones in coordinated 5G networks," in *Proceedings of Asilomar 2015*, 2015.